Cronfa
Swansea University's Research Repository

Rachel Bryan – Doctoral thesis re-submission, February 2025 – Swansea University

Swansea University
Prifysgol Abertawe

Linguistic Latent Conditions in International Nonproprietary Names of Medicines: Linguistic Drivers of Look-Alike Sound-Alike Errors and Implications in Russian and Italian Drug Nomenclatures

Rachel Bryan
BA (1st Hons), MA (Dist.)

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy

Swansea University
2024

1 Copyright: The Author, Rachel Bryan, 2025. Licensed under the terms of a Creative Commons Attribution-Non-Commercial No–Derivatives (CC-BY-NC-ND) license. Third party content is excluded for use under the license terms.

Abstract

This thesis has set out to identify linguistic latent conditions in international drug names. To achieve this aim, three research questions (RQ) are asked: *RQ1, How can the International Nonproprietary Name (INN) nomenclature be characterised in linguistic terms?*; *RQ2, To what extent are any of the linguistic features described in answer to RQ1 associated with the occurrence of Look-Alike Sound-Alike (LASA) error?*; and RQ3, *How can Russian-language and Italian-language pharmaceutical nomenclature be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?*

LASA errors are a type of medication error between similar drug names and constitute a significant threat to patient safety, e.g. hydralazine and hydroxyzine. They are highlighted in the World Health Organization's 2017 global patient safety challenge *Medication without harm.* The thesis frames linguistic features that increase the risk of LASA errors as a type of latent condition in the healthcare system that creates 'accident opportunities', under Reason's human error theory.

Three shortcomings in the literature are addressed: morphosemantic features of drug nomenclature (most studies look at formal features such as name length), linguistic properties of generic names (most look at brand names), and drug nomenclature in languages other than English. One linguistic latent condition is identified - initial trigram frequency – which leads to a suggestion for manufacturers and regulators to consider encouraging diversity in the beginning strings of names. Several other linguistic variables show potential associations with LASA error that merit further research but fall short of sufficient empirical evidence to make a policy recommendation. Two non-linguistic confounders are identified that significantly drive risk: prescribing frequency and year of approval. Analyses are complemented by two case studies of Russian-language INNs and Italian Common Names.

Declarations and statements

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ....Rachel Bryan.....
(candidate)
Date ....27 March 2024.....

Statement

This thesis is the result of my own investigations, except where otherwise stated. Sources are acknowledged by footnotes giving explicit references. A bibliography is appended. I hereby give consent for my thesis, if accepted, to be available for electronic sharing after expiry of a one-year bar on access approved by Swansea University.

Signed ...Rachel Bryan.....
(candidate)
Date ...27 March 2024.....

The University's ethical procedures have been followed and, where appropriate, ethical approval has been granted.

Signed .....Rachel Bryan.....
(candidate)
Date .....27 March 2024.....

Authorship statement

I hereby declare that all work in Chapters 1, 2, 4, 5, 6, and 7 are the sole work of Rachel Bryan (RB), with supervisory input from Dr. Jeffrey Aronson, Dr. Owen Bodger, Dr. Sue Jordan, Professor Pius ten Hacken, Dr. Alison Williams, and Dr. Giovanna Donzelli.

Table of contents

Abstract 2
Declarations and statements 3
Authorship statement 4
Table of contents 5
Acknowledgements 7
List of tables 9
List of figures 11
List of abbreviations and initialisms 12
Notation in the thesis 13
Chapter 1: Introduction 14
1.1 Background to the study 14
1.2 Research gaps 16
1.3 Research aims, objectives, and questions 18
1.4 Significance of the study 20
1.5 Structure of the thesis 20
Chapter 2: Background 22
2.1 INNs: an overview 22
2.2 INNs through a linguistic lens 26
2.3 Terminology usage in this thesis 32
2.4 The INN nomenclature as a taxonomy, and a typology 33
2.5 Conclusion 41
Chapter 3: Literature review 42
3.1 Introduction and background 42
3.2 Background to the problem of LASA errors 45
3.3 Theoretical aspects of LASA errors 58
3.4 Extrinsic and linguistic features of LASA name pairs 66
3.5 Framework of the field of LASA errors 70
3.6 Summary and Conclusion 80
Chapter 4: A taxonomic database of non-proprietary drug names in English 83
4.1 Introduction 83
4.2 Primary data collection 83
4.3 Orthographic features 90
4.4 Morphosemantic features 99
4.5 Relational features 116
4.6 Discussion, conclusion, and limitations 134
Chapter 5: Predicting the occurrence of LASA errors in English 137
5.1 Introduction 137
5.2 Methodology 138
5.3 Results 150
5.4 Discussion and conclusions 160
Chapter 6: Pharmaceutical nomenclatures in Russian and Italian 165
6.1 Introduction 165
6.2 Data analysis methods for RQ3: Multilingual implications 166
6.3 Case study 1: Russian-language INNs 168
6.4 Case study 2: Italian-language names (*Denominazioni Comuni Italiane*, *DCIt*) 188
6.5 Discussion and contrastive overview 208
6.6 Implications of findings from RQ2 in Russian and Italian 212
Chapter 7: Discussion and conclusions 216
7.1 Outline of research aim and questions 216
7.2 Research questions 1 and 2: Findings 217
7.3 Research Question 3: Findings 237
7.4 Conclusion 242
Appendix 253
A1 Glossary of terms 253
A2 Searches undertaken for narrative review of literature in health science, and case reports 254

A3 Searches undertaken for systematic review (Bryan et al., 2020a) 257
A4 Summary of medication error prevalence 262
A5 Communications log 263
A6 List of confused error pairs used as outcome measure in statistical analysis 265
A7 MedNet listing of salbutamol and its translations 275
A8 Methods to analyse formal measures 276
A9 Orthographic measures, over time 277
A10 INNs using a variant STEM 279
A11 Morphosemantic measures, over time 280
A12 Relational measures, over time 283
A13 Examples of ATC codes for names in four taxa 289
A14 STEMS used in INNs with a high number of neighbours 290
A15 All STEM taxa in English, Russian, and Italian nomenclatures 291
A16 Frequent STEM taxa in English and Russian 324
A17 Top ten beginnings of INNs in Russian and English (initial bigrams and trigrams) 327
A18 Russian and English initial bigrams and trigrams, descriptive statistics 327
A19 Top ten endings of INNs in Russian and English (terminal bigrams and trigrams) 328
A20 Russian and English terminal bigrams and trigrams, descriptive statistics 328
A21 Neighbourhood density in Russian and English 328
A22 Frequent STEM taxa in English and Italian 330
A23 Top ten beginnings of INNs in Italian and English (initial bigrams and trigrams) 333
A24 Italian and English initial bigrams and trigrams, descriptive statistics 333
A25 Top ten endings of INNs in Russian and English (terminal bigrams and trigrams) 334
A26 Italian and English terminal bigrams and trigrams, descriptive statistics 334
A27 Neighbourhood density in Italian and English 335
References 335

Acknowledgements

People often ask me how I came to study drug names, and I usually tell them that I love long words (and this is true – my mum remembers me reciting all the generic names of my eczema creams to my doctor when I was three!). But the academic link came in an undergraduate module where I created a termbase of five related terms, of my choosing. I decided to include adrenaline, since I have severe allergies and carry an Epipen. *Adrenaline* is the UK and EU name for this drug (with cognates such as *adrenalina*) but *epinephrine* is the US name and is used in Russia and many other nations. The reason for this difference is that when the EU harmonised their pharmaceutical nomenclature to the form of the International Nonproprietary Name, they deemed adrenaline too risky to change, as it is a lifesaving medication used in high-stress situations by healthcare practitioners, patients, and their families. I found the infrastructure of drug names utterly fascinating, ticking all the boxes for things that I was interested in - medicine, linguistics, languages, and long words! I went on to get Masters funding and then PhD funding to explore the topic in more detail.

This PhD has been a long journey, spliced up by an extended maternity leave and a very unproductive pandemic period. I am grateful to my family for their support in my endeavours, in particular Ben, for his loving support and providing me with the security to focus on the thesis, and for encouraging me during all the low moments and always believing in me. To Noah and Amy, my brilliant and inspiring children, for being a much-needed distraction from work and reminding me that I am a whole person outside of this thesis. Being a parent has punctuated and shaped my academic journey and has pushed me to work hard and be successful. To Belle, my melancholic dog, for all the deep-thinking dog walks. To my mum, Susanna, for reading the thesis and for all the willing babysitting and constant encouragement. Finally, to Michela Finnis and Sophie Hodgson for the Wales Coast Path walks that have cleared my mind, strengthened my body, and kept me level-headed. A special mention goes to Octonauts for entertaining my toddler while I wrote my literature review during the pandemic with no childcare.

I am grateful to Suzi Garcia for teaching me Italian in A Level college and not giving up on me when I fell pregnant halfway through the year, instead encouraging me to complete the course early and supporting me with UCAS applications and childcare grants. I wish I could thank her in person. I am grateful to Dr David Martin and Taisia Martin for teaching me Russian, bringing the language to life and instilling in me a deep love of the complexities of language.

For help during the PhD, I am grateful to a number of people. To Dr Alison Williams, my primary supervisor, for always being available and caring and especially understanding when I returned from maternity leave during my third year. To Dr Owen Bodger, my second supervisor, for extensive support with the statistical analysis (Dr Sue Jordan was second supervisor between 2016 and 2022), and Dr Giovanna Donzelli for assisting with the corrections. To Dr Jeffrey Aronson for invaluable feedback on clinical pharmacology, language, and internal processes of the INN Expert Committee. To Dr Pius ten Hacken for feedback on morphology and Russian and Italian linguistics, and for teaching me the foundations of linguistics as an undergraduate. To Professor Mary Gagen for strategy sessions and helping with the big picture thinking. To Rachel Tudor for help with some coding. And to my dad, Terry Bryan, for advice on healthcare regulation and patient rights.

It has been an honour to spend part of my life purely being inquisitive and exploring a topic in such detail and I am excited about future research paths. If you are about to read this thesis, thank you for your time and I hope you enjoy reading.

List of tables

Chapter 1: Introduction 14
Chapter 2: Background 22
Table 2.1 INN naming guidelines (WHO, 2017a) 24
Table 2.2 Comparing classifications of selegiline in INN and ATC systems* 34
Table 2.3 The -ast antiasthmatics taxon 36
Chapter 3: Literature review 42
Table 3.1 Examples of LASA name errors 51
Table 3.2 Usage of terms LASA and Look-Alike Sound-Alike 52
Table 3.3 Prevalence of LASA errors: a summary of prevalence noted or cited in the literature 53
Table 3.4 Rejected INNs 73
Table 3.5 Medication process, adapted from Table 1 in Lisby et al. (2010) 75
Chapter 4: A taxonomic database of non-proprietary drug names in English 83
Table 4.1 Primary data collection 84
Table 4.2 Summary of the distribution of orthographic variables in the taxonomic database 93
Table 4.3 Descriptive statistics for INNs (n=1,046), Orth-01 and Orth-02 94
Table 4.4 Prohibited letters and groups of letters in INNs, Orth-03 97
Table 4.5 Descriptive statistics for one-word INNs (n=1,046) 98
Table 4.6 Summary of the distribution of morphosemantic variables in the taxonomic database 103
Table 4.7 Common incidental STEMS in INNs, with examples 108
Table 4.8 Examples of FREEFIXES in INNs, with examples 110
Table 4.9 Headedness in INNS 110
Table 4.10 SUBSTEMS, agglutination and replacement 112
Table 4.11 Number of INNs in each taxon type 113
Table 4.12 Examples of pairwise scoring using the Trigram-2b measure 120
Table 4.13 Summary of the distribution of relational variables in the taxonomic database 123
Table 4.14 Top ten beginnings of INNs, compared with those of USANs (initial bigrams and trigrams) 125
Table 4.15 Top ten endings of INNs, compared with those of USANs (terminal bigrams and trigrams) 127
Chapter 5: Predicting the occurrence of LASA errors in English 137
Table 5.1 Pairwise and combined model of four confounders and the outcome measure, using Binary Logistic Regression 141
Table 5.2 Building the base model of three confounders (best performing highlighted in green) 143
Table 5.3 Variables to keep or exclude 146
Table 5.4 Collinearity between variables in the dataset: Pearson scores over 0.6. 147
Table 5.5 Collinearity: tolerance and VIF scores 147
Table 5.6 Variables in the dataset for statistical analysis 149
Table 5.7 Global model 150
Table 5.8 Backward elimination model, steps 151
Table 5.9 Final model 152
Table 5.10 Confusion matrix for final model 155
Table 5.11 Top twenty models in the bootstrap resamples 156
Table 5.12 Bootstrapping of final model 158
Table 5.13 Pairwise inclusion frequency 159
Chapter 6: Pharmaceutical nomenclatures in Russian and Italian 165
Table 6.1 Linguistic features and confounders that can be explored in Russian and Italian 167
Table 6.2 Names of top 50 most frequently prescribed medications in the UK and their Russian counterparts 171
Table 6.3 Orthographic features of Russian- and English-language INNs: length and Scrabble score 175
Table 6.4 Examples of STEMS in English and Russian that are formed by abbreviating the name of the pharmacological group 182
Table 6.5 Examples of STEMS in English and Russian that are formed by abbreviating the chemical name 183
Table 6.6 Initial and terminal trigrams in Russian and English (top ten most frequent, in descending order of frequency) 186

Table 6.7 Names of top 50 most frequently prescribed medications in the UK and their Italian counterparts _____ 191
Table 6.8 Orthographic features of Italian and English names: length and Scrabble score_____ 194
Table 6.9 Examples of STEMS in English and Italian that are formed by abbreviating the name of the pharmacological group _____ 202
Table 6.10 Examples of STEMS in English and Italian that are formed by abbreviating the chemical name _____ 203
Table 6.11 Initial and terminal trigrams in Italian and English (top ten most frequent, in descending order of frequency) _____ 206
Chapter 7: Discussion and conclusions 216
Table 7.1 Summary of findings on orthographic features of INNs _____ 219
Table 7.2 Summary of findings on morphosemantic features of INNs _____ 224
Table 7.3 Summary of findings on relational features of INNs _____ 229
Appendix 253
References 335

List of figures

Chapter 1: Introduction 14
Figure 1.1 Mind map of Research Questions 20
Chapter 2: Background 22
Figure 2.1 Nomenclature of salbutamol 23
Figure 2.2 Type 1: Single-level taxon -coxib, for 'selective [COX2] cyclo-oxygenase inhibitors', and example parecoxib 37
Figure 2.3 Type 2: Multi-level taxon -ast, for 'antiasthmatic, antiallergic substances not acting primarily as antihistaminics', with examples montelukast and roflumilast 37
Figure 2.4 Type 3: Complex taxon -mab, for 'monoclonal antibody' and example trastuzumab; old system up to April 2017 39
Figure 2.5 Type 4: Irregular taxon vir, for antivirals 40
Figure 2.6 Type 4: taxon vir, for 'antivirals' and example aciclovir 40
Figure 2.7 Type 4: taxon vir, for 'antivirals' and example maraviroc 41
Chapter 3: Literature review 42
Figure 3.1 Search methodology for Chapter 3 43
Figure 3.2 Swiss cheese model (SCM) - attribution David Mack / CC BY-SA 60
Figure 3.3 Hot cheese model, reproducing Figure 2 from Li & Thimbleby (2014) 61
Figure 3.4 Relationship between error and harm: a Venn diagram 65
Figure 3.5 Conceptual framework for Chapter 3 71
Chapter 4: A taxonomic database of non-proprietary drug names in English 83
Figure 4.1 Prescribing frequency, by decade 88
Figure 4.2 Density plot for Orth-01, number of characters 95
Figure 4.3 Density plot for Ortho-02, number of syllables 95
Figure 4.4 Bar chart for Orth-04, presence of prohibited letters 97
Figure 4.5 Density plot for Scrabble score 98
Figure 4.6 Density plot for Orth-4, normalised Scrabble score 98
Figure 4.7 Bar chart for Morph-01, compliance with WHO guidelines 105
Figure 4.8, Bar chart for Morph-02 106
Figure 4.9 Bar chart for Morph-03 106
Figure 4.10 Bar chart for Morph-04 107
Figure 4.11 Density plot for Morph-05. 108
Figure 4.12 Bar chart for Morph-06 109
Figure 4.13 Bar chart for Morph-07 111
Figure 4.14 Bar chart for Morph-08 113
Figure 4.15 Density plot for Rel-01, initial bigrams 126
Figure 4.16 Density plot for Rel-02, initial trigrams 126
Figure 4.17 Density plot for Rel-03, terminal bigram 128
Figure 4.18 Density plot for Rel-04, terminal trigram 128
Figure 4.19 Density plot for Rel-05, STEM frequency 129
Figure 4.20 Density plot for Rel-06, broad alignment with ATC index 130
Figure 4.21 Density plot for Rel-07, specific alignment with ATC index 131
Figure 4.22 Density plot for neighbourhood density 132
Figure 4.23 Density plot for Rel-08, neighbourhood density 132
Figure 4.24 Bar chart for Rel-09, neighbourhood frequency 133
Figure 4.25 Density plot for Rel-10, semantic transparency 134
Chapter 5: Predicting the occurrence of LASA errors in English 137
Figure 5.1 Prescribing frequency and LASA error 142
Figure 5.2 Decade and LASA error 142
Figure 5.3 Prescribing frequency and decade 142
Figure 5.4 Morph-09 STEM taxon, and error 145
Figure 5.5 Parameter estimates of variables in final model 153
Figure 5.6 Groups in Morph-09, STEM taxon 154
Chapter 6: Pharmaceutical nomenclatures in Russian and Italian 165
Figure 6.1 Paracetamol listing 169
Figure 6.2 Hyperonymic STEM building in montelukast (reproduced from Figure 2.3) 180
Figure 6.3 Distribution of neighbourhood density in Russian 188
Figure 6.4 Paracetamol listing (reproduced from Figure 6.1) 189

Figure 6.5 Hyperonymic STEM building in montelukast (reproduced from Figure 2.3) _____ 200
Figure 6.6 Distribution of neighbourhood density in Italian _____ 208
Figure 6.7 Distribution of initial trigrams in Russian and Italian _____ 210
Figure 6.8 Distribution of terminal trigrams in Russian and Italian _____ 211
Chapter 7: Discussion and conclusions 216
Appendix 253
References 335

List of abbreviations and initialisms

Full term	Abbreviation or Initialism
Adverse drug reaction	ADR
Agenzia Italiana del Farmaco (Italian Pharmacy Agency)	AIFA
Anatomical Therapeutic Chemical classification	ATC
British Approved Name	BAN
Barcode medication administration	BCMA
British Pharmacopoeia Commission	BPC
Clinical decision support	CDS
Computerised physician order entry	CPOE
Denominazione Comune Italiana (Italian Common Name)	DCIt
Exploratory data analysis	EDA
European Medicines Agency	EMA
Food and Drug Administration	FDA
Global Patient Safety Challenge	GPSC
Hot cheese model	HCM
Healthcare practitioner	HCP
Human Genome Organization [gene nomenclature]	HUGO
International Nonproprietary Name	INN
Institute of Safe Medication Practices	ISMP
Joint Commission on Accreditation of Healthcare Organizations	JCAHO
Look-Alike Sound-Alike [errors]	LASA
Look-Alike Sound-Alike, Read-Alike [errors]	LASARA
Longest common subsequence ratio	LCSR
Language for general purposes	LGP
Language for special purposes	LSP
Medication errors reporting programme	MERP
Medical Subject Headings	MeSH
National Health Service	NHS
Positional family size [ratio]	PFS
Phonetic and Orthographic Computer Analysis	POCA
Research question	RQ

Sound-Alike, Look-Alike [errors]	SALA
Swiss cheese model	SCM
Semantic transparency	ST
Farmacopea Ufficiale Italiana	FUI
United States Adopted Name	USAN
University of Washington Digital Anatomist [project]	UWDA
Vaccine errors reporting programme	VERP
World Health Organization	WHO

Notation in the thesis

Throughout this document, INNs are in blue and official STEMS are in green. When a STEM is the focus of discussion, it may be highlighted in the name, e.g. tamoxifen. Position of the STEM is shown by a hyphen; FREEFIX STEMS are not given a hyphen. Russian-language INNs are in purple and Italian Common Names are in orange. Phonemes are given in /forward slashes/, graphemes are given in <angle brackets>, foreign-language words are in *italics*, and glosses of foreign words are given in ‘single quotes’. Nonce words, including hypothetical medication names, are given an *asterisk. Linguistic terms that are applied a specific pharmacological usage are in SMALL CAPS.

Chapter 1: Introduction

This chapter presents the background to the study, the research problem, and the research aims and objectives. It introduces the three research questions and the significance of the study. Finally, it briefly outlines the structure of the thesis.

1.1 Background to the study

In 2017 the World Health Organization (WHO) launched their third Global Patient Safety Challenge (GPSC), *Medication without Harm*, asserting that all medication errors are potentially avoidable and can be reduced or prevented by improving systems and practices (WHO, 2017b). The challenge aimed to reduce the level of severe, avoidable harm related to medications by 50% over 5 years (i.e. by 2022) globally. The WHO has worked to forward the aims of the GPSC by stimulating and supporting national patient safety initiatives, blending evidence-based interventions with multi-modal implementation strategies; the work has been designed to gather momentum like a social movement.

The global associated annual cost of medication-related harms is $42 billion USD (WHO, 2017b). In addition to the moral and economic imperatives, there is a social imperative to reduce medication-related harms, with low income countries losing twice as many disability-adjusted life years as a consequence of adverse events as high income countries (WHO, 2017b). The GPSC is embedded in the WHO's patient safety philosophy that errors are broadly provoked by weak health care systems, and so the challenge aims to reduce their incidence, prevalence, and impact. The WHO concedes that health care systems are complex and often dysfunctional, but can be made more resilient to risk and harm if they are better understood through effective research and development:

> *Experience from other high risk industries … demonstrates that ... human beings make mistakes rarely through neglect, but instead because the systems, processes and procedures that they work with are often flawed or dysfunctional. This inevitably gives rise to errors and medication harm is no exception to this rule (WHO, 2017b: 4).*

Pharmacovigilance and medication safety are not new activities. In 2002 the WHO cited new patient safety concerns born from globalisation, consumerism, growth in free markets, and increased cross-border communication (WHO, 2002). Furthermore, the increasingly global

trade in pharmaceuticals and higher levels of regulatory complexity have prompted many intergovernmental bodies to make efforts towards harmonisation of regulatory activities, in order to ensure consistent efficacy of pharmacovigilance efforts (Bryan, 2015; WHO, 2002, 2017a). In the UK, the National Health Service (NHS) has consistently outlined medication errors causing serious harm as an 'area for improvement' (Care Quality Commission, 2023; NHS, 2023).

In the UK, of all events that are reported to cause patient harm (including but not limited to 'serious patient harm'), medication errors are the most common. Between April 2017 and March 2018 alone, they accounted for 11% of incidents,[1] and 57 deaths (NHS, 2018a, 2018b). It is estimated that medication errors cost the USA between $15 and $28bn each year, and that the USA spent an additional $213bn (8% of total healthcare spend) in 2012 on costs arising from medicines mismanagement, including medication errors (Aitken et al., 2013).

Medication errors can occur when medications have similar-looking or similar-sounding names, such as hydroxyzine and hydralazine, and/or shared features of product packaging. These 'wrong drug' errors (under the 5Rs framework) are so-called *Look-Alike Sound-Alike* (LASA) errors (Bryan et al., 2015). The five rights, or '5Rs' colloquially, emphasise the importance of giving the right drug, to the right patient, at the right dosage, via the right route, at the right time (Bundy et al., 2009)[2]. LASA errors make up a high proportion of all medication errors; estimates of prevalence range from: 2 in 1,000 to 3 in 100,000 prescriptions, 7% of near misses, and between 6.2% and 22% of all medication error events. They thus represent a significant threat to patient safety (Aronson, 2009a; Emmerton & Rizk, 2012; Lambert et al., 1999; Ostini et al., 2012; Runciman et al., 2003). They may occur during prescribing, dispensing, or administration of medicines and may result in overdosing, under-dosing, or inappropriate dosing (Aronson, 2009a). Confusion can occur between: generic-generic names (e.g. lorazepam - alprazolam); brand-brand names (e.g. Evista - Avinza); brand-generic names in either direction (e.g. Soriatane confused for sertraline, or methadone - Metadate) - these examples are taken from error reports listed in ISMP (2019). Most LASA pairs are reciprocal, i.e. each has been mistaken for its counterpart, revealing

[1] 204,162 medication incidents out of a total of 1,942,179 incidents

[2] To note, the National Institute of Clinical Excellence has proposed a sixth right, the 'right to decline' (NICE, 2024)

the influence of inherent pairwise similarity, rather than external environmental factors (Bryan et al., 2020a). Key terminology is defined in the glossary in Appendix A1.

'Naming, labelling and packaging' of medicines is one element of the 16-part strategic framework of the GPSC, as the WHO acknowledges:

> *Medicines are sometimes complex and can be puzzling in their names, or packaging, and sometimes lack sufficient or clear information. Confusing 'look alike sound alike' medicine names and/or labelling and packaging are frequent sources of error and medication-related harm that can be addressed (WHO, 2017b).*

LASA errors occur because of shared linguistic properties between two or more names, and lead to incorrect substitution of one medication for another. An example of a LASA error between two generic names, and the aetiology of the error, is given in below:

> Intended: **mercaptamine**; confused: **mercaptopurine**. In 2010 in the UK, the MHRA reported that an infant with nephropathic cystinosis was given mercaptopurine instead of mercaptamine. After taking the wrong medication for one month, the infant developed pancytopenia but made a full recovery after the error was noticed and rectified (MHPRA, 2010).

If not interrupted, this results in erroneous drug administration. The cause of the error may lie in similarities in orthography (the written forms) or phonology (the spoken forms), and several similarity measures have been proposed (Kovacic & Chambers, 2011). The potential for LASA errors may also be compounded by similarities in product packaging, dosage form or strength, or therapeutic indication. LASA errors often involve selection of the wrong medication from a shelf or from an electronic list (Emmerton & Rizk, 2012).

1.2 Research gaps

LASA errors are an emerging problem in the patient safety literature, and there are numerous studies exploring their aetiology, impact, prevalence, and economic burden. An overview is given in Bryan et al. (2020a), complemented by a systematic review in Bryan et al. (2020b). Studies tend to fall on one of three axes: probability of error, opportunity of error, and seriousness of error. This thesis is looking for features that increase the probability of error. Various solutions to error are presented in the literature and can be implemented:

during the pre-approval stage (preventing future errors), during the treatment process (avoiding current errors), or learning and reporting (recalling past errors). All three stages are important; however, there is growing alignment in the literature towards pre-approval solutions, driven heavily by Bruce Lambert and colleagues (e.g. Lambert et al., 2005).

Approaches to error causation and reduction fall into two camps, aligning with James Reason's person and system approaches (Reason, 2000). There is general consensus in the literature to prefer the system approach, viewing humans as fallible and presenting error as caused by latent conditions in the system that must be identified and safeguarded against (Emmerton & Rizk, 2012; Reason, 2000; Tuohy & Paparella, 2005). There are many latent conditions that increase the probability of LASA errors, and these are both linguistic (inherent) features of the names, and extrinsic features, such as whether names appear together on a shelf or in a drop down list.

This thesis addresses three gaps in the literature:

1. A lack of studies of semantic features of nomenclature (most look at formal features)
2. A lack of studies of the linguistic properties of generic names (most look at brand names)
3. A lack of studies on pharmaceutical nomenclature in other languages

First, studies focus almost exclusively on formal aspects of drug names (Lambert et al., 2001a), especially orthographic and phonological [3]similarities between names. The lack of focus on semantic properties in the INN nomenclature should be addressed. An extensive STEM system to communicate meaning has been created by the WHO, and its use is strongly encouraged. It is important to explore whether it is being used predictably and to the benefit of its users. Previous work found that names with high levels of orthographic similarity were more likely to share semantic features, i.e. a STEM, see Bryan et al. (2015). Other studies have found that semantic similarity enhances item recall (Lambert, Chang, & Lin, 2003). There is thus complex interplay in the nomenclature between semantic and formal aspects, which warrants closer analysis.

[3] A note on terminology: *phonological* refers to the codified system of pronunciation and features associated with names as they are spoken and heard, such as those found in a dictionary, while *phonetic* refers specifically to physical speech signals; although organisations such as the WHO often use these terms interchangeably (e.g. the Phonetic and Orthographic Computer Analysis tool), *phonological* is used here to describe standardised pronunciation and *phonetic* for the specific acoustic signals of speech.

Secondly, there is a concentration of studies on LASA errors between brand names (e.g. Basco et al., 2010; Bulfone et al., 2012; Lambert et al., 2005; Shah et al., 2017). This probably reflects the proportion of LASA errors between brand names to those between generic names (based on pairs listed by ISMP (ISMP, 2019), 58% versus 33% respectively). Studies on generic names explore national naming systems, such as USANs or BANs (Lambert et al., 2001a; Zhong et al., 2016). INNs are designated by the WHO, published in seven languages, and formally placed in the public domain and encouraged to be used internationally as the default form of the generic name. Many healthcare systems stipulate that prescribers use the generic name to avoid using costly medications rather than cheaper generic alternatives. The lower proportion of LASA errors between generic names may indicate that they are less likely to be confused; however, since their use is strongly encouraged in the UK, it will probably increase and so their linguistic properties warrant deeper exploration.

Thirdly, there are very few studies of drug nomenclatures in languages other than English. Literature searches are restricted to publications written in English, Italian, and Russian, and the analytical components of the thesis look at nomenclatures in these languages. Findings from any analyses in this field should be considered within a broader international context, because drug names are used in languages other than English and are translated and transliterated into other languages, alphabets, and writing systems. English is chosen as a focal language for examination, owing to its pivotal role in the WHO INN programme and because in general it remains the international language of science. It serves as the language of INN naming guidance and the primary operating language for the INN programme and is considered the 'source' language from which all other INNs and generic names are translated. Additionally, Italian and Russian are selected as case-study languages, each representing unique points of interest. Italian is a Romance language that uses the Roman alphabet, with a national naming system outside the WHO nomenclature. In contrast, Russian is a Slavic language using the Cyrillic alphabet, with INNs designated by the WHO, because Russian is one of the six official languages of the United Nations.

1.3 Research aims, objectives, and questions

The research aim is to identify linguistic latent conditions in international drug names – linguistic properties in the system of medication names that drive the risk of LASA errors, and to consider them in the context of Russian- and Italian-language drug nomenclatures.

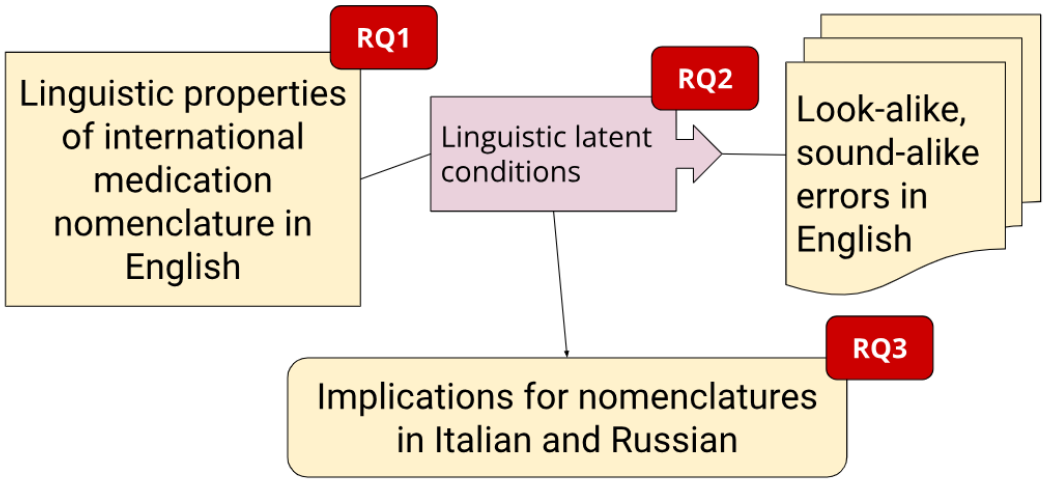
The research objectives are threefold: first, to explore linguistic features of the INN nomenclature; second, to explore the extent to which these features are associated with the occurrence of LASA errors; and third, to consider how these findings could be extrapolated to nomenclatures in other languages and writing systems.

These objectives translate into the following three research questions:
RQ1. How can the INN nomenclature be characterised in linguistic terms?
RQ2. To what extent are any of the linguistic features described in answer to RQ1 associated with the occurrence of LASA error?
RQ3. How can Russian-language and Italian-language drug nomenclatures be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?

These three research questions are closely related and are co-dependent. Once linguistic properties have been described (fulfilling RQ1), they will be analysed statistically for correlation with LASA errors (fulfilling RQ2) to identify linguistic latent conditions. The implications of these linguistic latent conditions to names in other languages will then be considered (fulfilling RQ3). All three research questions are novel.

Look-Alike Sound-Alike is a property that can be applied to medication name pairs (a *LASA pair of names*) as well as medication errors (a *LASA error*). In this study both are in focus because both the potential for error as well as the actual occurrences of error are of interest. The focus is on generic medication names, specifically INNs designated by the WHO in English and Russian, and names designated in Italian by the *Farmacopea Ufficiale Italiana* (FUI, 'Official Pharmacopoeia of Italy'). It is focussed on LASA medication names and errors that are linguistic; i.e. due to similar names, not similar packaging or look-alike labelling alone. Both orthographic and phonological similarities will be taken into account. Look-alike unit errors (such as 2iu/21u) are excluded since they do not relate to the linguistic form of the medication name but rather the individual letter glyphs. Despite designation of a medication as an INN, it may never be authorised for use in any country, simply being 'reserved' for that substance. Names for medication that are authorised in the UK are within the scope of this study since it looks at the likelihood and risk of error. At any rate, the risk that an unauthorised medication will be involved in a LASA error is extremely low because its use is restricted to clinical trials until authorisation. RQ1 (linguistic analysis) and RQ2 (correlation with LASA errors) are explored for the English nomenclature only, because naming guidance from the WHO is given solely for the English nomenclature, and published lists and case reports of LASA error are available only in English.

Figure 1.1 Mind map of Research Questions



1.4 Significance of the study

The thesis works to enrich the knowledge base of medication nomenclature by exploring linguistic properties of names. Findings can inform the pre-approval stage by identifying linguistic properties in proposed names that should be avoided and recommending quantitative measures to use at this stage of the process. Findings may go some way to identify or suggest potential LASA name pairs (involving names that have already been approved) that should be added to watch lists.

Very little attention is given to the linguistic properties of medication names, beyond studies measuring and predicting orthographic or phonological similarities between names (Lambert, Chang, & Gupta, 2003; Lambert et al., 2001a). This thesis addresses that gap in knowledge, by conducting a broad analysis of linguistic properties of medication names to tease out properties that correlate with the risks of LASA errors. These linguistic properties are framed as 'linguistic latent conditions' within Reason's system approach (Reason, 2000), and are presented as potential risks to consider during the naming process.

1.5 Structure of the thesis

The thesis has seven chapters. Chapter 1 is the introduction and gives an overview of the study and presents the research aims and objectives. Chapter 2 summarises relevant theory

from linguistics. Chapter 3 reviews the literature on LASA errors, and is complemented by two papers published in the *British Journal of Clinical Pharmacology* in 2020 and 2021 (Bryan et al., 2020a, 2020b).

Chapter 4 describes the creation of a taxonomic database of INNs, presenting descriptive statistics to answer research question 1. Chapter 5 explores association between linguistic properties and the occurrence of LASA errors, presenting inferential statistics to answer research question 2. Chapter 6 considers the multilingual implications of relevant linguistic latent conditions for nomenclatures in other languages, using Italian and Russian as example case studies, to answer research question 3. Chapter 7 presents the discussion and conclusions, considering how the three research questions have been answered, as well as limitations and avenues for further research.

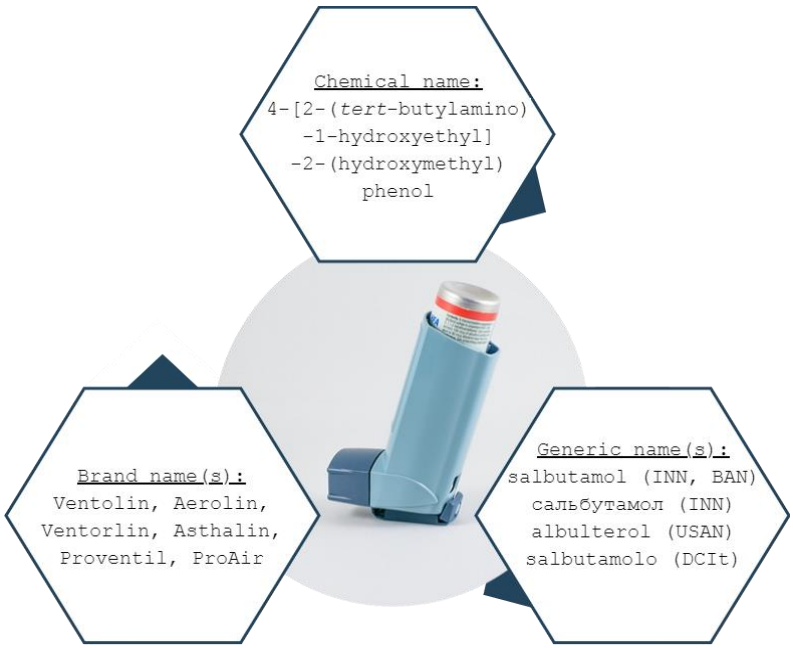
Chapter 2: Background

This is the background chapter, and briefly comprises an overview of International Nonproprietary Names (INNs), key concepts in linguistics that are relevant to the study, a discussion of terminology usage and how it differs between the fields of linguistics and pharmacology, and a typology of INN taxa, first presented in Bryan (2015).

2.1 INNs: an overview

Before new medicines can be brought to market, their names and designations (their 'terminology') must be agreed. In pharmacology in particular, medications are named within a highly complex nomenclature system overseen by multiple governmental organisations, illustrated in Figure 2.1 using the example of salbutamol. A medicinal substance is assigned a minimum of two names: a chemical name and a generic name. If they reach the marketing and authorisation stage they will be assigned a proprietary 'brand' name (see Figure 2.1). The chemical name is designated by the International Union of Pure and Applied Chemistry (IUPAC) and is based on the chemical formula of the substance, indicating the position of hydroxyl groups, the length of the carbon chain, and so on. A medication that is authorised for use will have at least one brand name, chosen by each manufacturer (e.g. Ventolin). The name is commercially motivated, initially capitalised, may be followed by the registered trademark symbol, and is legally required to imply no therapeutic benefit. It is typically laconic and euphonious. Once out of patent (up to 20 years in the EU), a substance can be marketed and sold by competitors and will be given more brand names. In addition, in each country that the medication is licensed to be marketed, it will be assigned a national generic name (such as a British Approved Name, BAN, in the UK, a United States Adopted Name, USAN, or a Denominazione Comune Italiana in Italy). It will also be assigned a global generic name - an International Nonproprietary Name (INN) by the World Health Organization (WHO). Other examples include paracetamol (INN, BAN), acetaminophen (USAN), paracetamolo (DCIt). In antiquity, medications were named haphazardly, often after the gods, such as *morphine* from Morpheus, the Greek god of dreams, or anandamide from Sanskrit *ananda*, 'bliss, delight' (Oxford English Dictionary [OED], 2022). Elements of Latin and Greek languages persist in word formation processes in English, in particular for terms in specialist domains such as medicine.

Figure 2.1 Nomenclature of salbutamol



There are multiple strata in the nomenclature of generic names: they are formed on international, regional, and national levels, and are published multilingually. International Nonproprietary Names (INNs) are the most commonly used system of generic names, and their form is used as the default national name in many systems, such as in the EU and UK (with BANs), with only a few notable exceptions (Aronson, 2000; EU Directive, 1992).

The WHO launched the INN programme in 1952, and to date have assigned around 9,000 INNs, around 120 per year. The names are formally placed in the public domain to promote consistency of global communication between manufacturers, clinicians, prescribers, carers, and patients. The INN nomenclature is published in the six official languages of the United Nations (English, French, Spanish, Arabic, Chinese, and Russian), and Latin (WHO, 2017a). The Latin form of the name is used as the source form for translation into other European languages, such as Italian and Portuguese (Marecková & Cerveny, 2002). An example of an INN in all its published languages is given in Figure A3.1. Many of the medications referred to are in the early stages of discovery and clinical trials, and may never receive marketing authorisation - only names referring to medications authorised in the UK are included within the scope of this study.

INNs are designated according to a set of primary and secondary guidelines (see Table 2.1). The naming guidelines stipulate spelling, phonology, hyphenation, word length, and semantic relationships. They aim to achieve:

- Usability: is the name usable in all four modalities of language - reading, writing, listening, and speaking? Is it pronounceable, legible, audibly perceptible, comprehensible, printable, and memorable?
- Taxonomy: how does the name show semantic interrelationship within the conceptual system (the taxonomy). This requires first that there is a robust and consistent conceptual system in place, and secondly that formal properties of language are exploited to map these semantic relationships onto the orthographic and phonological realisation of the name.
- Clarity: how liable is the name to be confused with other names, both in the same system and in other systems?
- International use: does the name adhere to the phonotactics of all languages in which it is used, and can it be easily transmitted into other languages with different alphabets or writing systems? (Bryan et al., 2015)

Table 2.1 INN naming guidelines (WHO, 2017a)

Number	Guideline
1	INNs should be distinctive in sound and spelling. They should not be inconveniently long and should not be liable to confusion with names in common use.
2	The INN for a substance belonging to a group of pharmacologically related substances should, where appropriate, show this relationship. Names that are likely to convey to a patient an anatomical, physiological, pathological or therapeutic suggestion should be avoided.
3	In devising the INN of the first substance in a new pharmacological group, consideration should be given to the possibility of devising suitable INN for related substances, belonging to the new group.
4	In devising INN for acids, one-word names are preferred; their salts should be named without modifying the acid name, e.g. "oxacillin" and "oxacillin sodium", "ibufenac" and "ibufenac sodium".
5	INN for substances which are used as salts should in general apply to the active base or the active acid. Names for different salts or esters of the same active substance should differ only in respect of the name of the inactive acid or the inactive base. For quaternary ammonium substances, the cation and anion should be named appropriately as separate components of a quaternary substance and not in the amine-salt style.
6	The use of an isolated letter or number should be avoided; hyphenated construction is also undesirable.
7	To facilitate the translation and pronunciation of INN, "f" should be used instead of "ph", "t" instead of "th", ”e” instead of “ae" or "oe", and “i" instead of “y"; the use of the letters “h" and 'k" should be avoided. When devising an INN it is important to be aware of possible language

	problems. Since the name is used worldwide, not only should certain letters be avoided, but experts need to be aware of unsuitable connotations in the major languages spoken in the world.
8	Provided that the names suggested are in accordance with these principles, names proposed by the person discovering or first developing and marketing a pharmaceutical preparation, or names already officially in use in any country, should receive preferential consideration.
9	Group relationship in INN (see Guiding Principle 2) should if possible be shown by using a common stem. [...]

INNs use an extensive system of STEMS[4] to communicate meaning and their use is strongly encouraged in naming guideline 2, in Table 2.1. These STEMS are compositional word parts used by the WHO to indicate pharmacological relationship (WHO, 2017a, 2021b). The use of STEMS reflects an underlying taxonomic conceptual system for INNs, and allows users to exploit this systematicity to improve retention, pronunciation, and recognition of the names. STEMS may be acronyms (e.g. -mab for *monoclonal antibody*, in trastuzumab) or abbreviations (such as -vir for *antivirals*, in aciclovir; or -ast for *antiasthmatics*, in montelukast). They are often indirectly formed from neoclassical compounds. For example, the SUFFIX STEM -kin is an abbreviation of *interleukin*, which in turn is a neoclassical combining form of the morphs inter- (Latin, 'between') and -leuk- (Greek λευκο-, 'white'), plus the suffix -in.

Since there is no expiration date for a published STEM, nor for a published INN (see guideline 3 which attempts to future-proof the nomenclature), the nomenclature is getting denser all the time and some of the rules will change gradually to account for this growth, e.g. the recent change prohibiting duplicate letters and the iterative changes to naming of monoclonal antibodies - "the INN Programme runs on a flexible basis with living documentation underpinning it" (Balocco et al., 2022; SoINN, 2020a).

Although a substance can have only one chemical name, as assigned by IUPAC, it may be assigned different generic names by each country in which it is approved, and it may be marketed under multiple brand names. The sheer number of names for substances and regional or dialectal variations increases the risk of confusion, and medical advances will lead to further proliferation of the pharmaceutical market and these inherent risks.

[4] In this thesis, SMALL CAPS are used to indicate terminology that is used with a pharmacological sense, for example STEM, SUBSTEM, AND AFFIX. This is described in full in Section 2.3.

2.2 INNs through a linguistic lens

In this section, INNs are grounded in the field of linguistics, introducing concepts and frameworks that are relevant to this study. Firstly, the linguistic first principles theory of the *linguistic sign* is discussed, and the centrality of morphology and semantics to the thesis. Secondly, INNs are couched in terms of them being part of a nomenclature, as quasi-neoclassical compounds, and as meaningful units with variable semantic transparency.

2.2.1 First principles: semantics, morphology, and Saussure's linguistic sign

Linguistics is traditionally divided into subfields (classification adapted from Bauer, 2007) relating to both formal (written or spoken form) and semantic aspects (meaning or conceptual structure). Semantics is highly relevant to this study, as it examines how meaning is constructed in language, both at the word level and within sentences. Medication names directly refer to the substances they represent, and understanding these names requires a focus on the relationships between the meanings of words. Semantics also links closely with morphology, which studies how words are structured. Medication names consist of meaningful components that are systematically combined during the naming process, making morphology central to understanding how these names are formed. Semantics and morphology will be discussed in more detail in this chapter, as well as highlighting morphological and semantic features to analyse.

The relevance of semantics and morphology reflects the first principles theory of Saussure's *linguistic sign*. Linguistic sign theory is a useful framework to understand basic aspects of *form* and *meaning*, and despite rapid developments in the field of linguistics since Saussure's time, it remains a useful didactic tool for representing the relationship between form and meaning in language. It is also a useful way to ground this novel study that applies linguistic analysis to a problem in health care.

In Saussure's theory, the linguistic sign comprises two inseparable and mutually co-occurring parts: the *signifier*, the acoustic or graphic actualisation of the word; and the *signified*, the physical imprint in the mind of the concept or referent in the real world that is recalled upon hearing or reading the *signifier*. These may be seen as two sides of a coin, each irrevocably linked to the other; one side can exist independently only in an abstract sense (Joseph, 2004). According to Saussure, the sum value of the linguistic sign is determined by the degree and manner in which it is different to other words. Jacques Derrida coined the theory of *différance*, a blend of difference and deference – no one linguistic sign

can encapsulate a meaning, a term is defined by what it is *not* and meaning is gathered by studying the 'associative relations' of the word. This aligns closely with Saussure's theory of language. "The sign ... is deferred presence", and "every sign rests purely on a negative co-status" (Derrida, 1982).

Like two sides of a coin, form and meaning (formal and semantic aspects of language) are inextricably linked: formal aspects of INNs are motivated by the semantics they represent, and the form (both written and spoken) of INNs can both facilitate and obfuscate meaning and contribute to the risk of confusion. Form and meaning will be used as the basis for analysis in this study, and the following sections give an overview of relevant linguistic principles related to form and meaning, and place these within the context of INN nomenclature and LASA errors where appropriate.

Form is typically taken to mean the spoken form of the name, but in the case of medication names their written form of INNs is an important aspect of their linguistic representation. In the first instance, the WHO publishes INNs in written format, and names are written in a range of contexts: via computerised order entry systems, handwritten and printed on prescriptions, and printed on labels and packaging. Although phonology is not within the scope of the study, the written form of INNs is easily analysable, and so orthographic features can be analysed. Orthography is not traditionally a field in linguistics, because more abstract notions are considered beyond letters and spelling. In the case of LASA errors, however, it is important to look at the more mechanical features, such as the letters that make up the names and how they are written, typed, printed, transcribed, and read. Furthermore, if INNs are found to be orthographically shallow (i.e. if their written form is closely mapped onto their phonological form), findings are likely to be extrapolable to sound-alike errors (the 'SA' in LASA).

2.2.2 INNs as a nomenclature

INNs are terms that form part of a nomenclature, a closed collection of names for concepts in a specific domain, in this case generic medication names. They have certain characteristics – they are all names and are formed according to a set of rules determined by an official body. There is a useful distinction to be made between words and terms. Since all terms are words (or multiple words), but not all words are terms, *term* is a subcategory of *word*. It follows that the definition for term should be narrower than that for word. ISO 1087 Vocabulary of Terminology (1087:5.3.1.2) defines a term as the "designation of a defined concept in a special language by a linguistic expression", in which 'designation' is "any

representation of a concept" and 'concept' is "a unit of thought constituted through abstraction on the basis of properties common to a set of objects". The use of 'special language' in this definition is of particular importance: one way in which words and terms can be distinguished is to study their usage in particular linguistic settings, and often language is distinguished between general (LGP) and special purposes (LSP). This distinction originates in Zellig Harris' sublanguage theory and proposes that certain contexts influence language use and place restrictions on vocabulary and syntax. Vocabulary in LSP is said to comprise terms, while vocabulary in LGP comprises words (Pearson 1998:12).

Cabré (1999:33) differentiates terminology from lexicography in that while lexicography views the lexical unit as central and meaning as a related component, terminology assumes that a concept can exist independently of a name. The traditional approach in terminology is onomasiological: it begins with a set of concepts and designates a name to each of them, asking "how do we express X?"; lexicography is semasiological: it begins with a set of words and describes their meaning, for example in a dictionary. Lexicology is always linked to grammar, whereas in terminology the term is central, and "neither inflection … nor syntax … are of consequence" (loc. cit.). Unlike linguistics and lexicography, terminology is usually synchronic and prescriptive, its primary purpose being the standardisation of terms for current use. Term formation is normally onomasiological, the concept being seen as the determining factor of designation, although the terminographical process resulting from term formation will have a semasiological facet, in that the terminologist will 'reverse engineer' the terms to reveal the underlying system of concepts.

Wüster is regarded as the forerunner of the emergence of terminology as a distinct branch of linguistics (Pearson 1998:10). He argued that the concept is central to the study of a term, while in the study of general language vocabulary, the name is central. Concepts exist independently of language and of their labels and should be viewed as mental constructs. Wüster was interested only in the naming of concepts and vocabulary, assigning morphology and syntax to the realms of general linguistics, and he argued that terms for concepts in particular subject fields needed to be standardised and used consistently to ensure smooth communication, with a one-to-one correspondence between label and concept. Pearson disagreed with Wüster's approach, arguing that to disregard morphology and syntax was to oversimplify the role of terms in a language, and to view terms as simple labels for concepts was to relegate them to the level of proper names.

The lexicon of a special language may be seen as a reflection of the system of concepts in the related subject domain. Lexical units are provided for each concept that is

"conventionally established in the subspace" and their reference is restricted to one "well-defined region" (Sager, 1990). The lexicon of a special language also contains "items of general reference", which are not specific to any discipline and have "vague or generalised" denotations. These are 'words' and they are used in a "variety of sublanguages"; the items "characterised by special reference" are 'terms'. The general language exploits stylistic devices such as metaphor and idiom, and permits fuzziness in meaning delimitation, leading to polysemy. Special languages "strive to systematise principles of designation and to name concepts according to pre-specified rules or general principles" (Sager, 1990): 57). Ten Hacken distinguishes between words and terms by their origin, in that a term is the result of explicit teaching, but a word is acquired naturally (Ten Hacken, 2010).

2.2.3 INNs as quasi-neoclassical compounds

Neoclassical word formation is a process in which a morph (a word part) is borrowed from a classical language and combined with other morphs to form a word with a new meaning that does not exist in the classical language. An analogy would be using parts from a classic car to create a new, modern car. This is helpful to users of the specialist language, as the terms formed are likely to be understood by speakers of various European languages, because they share lexical roots. For example, *neuroblastoma* comprises neuro (Greek νευρο-, 'nerve'), blast (Greek βλαστός, 'embryo') oma (Greek ωμα, 'tumour'). In English, morphemes can be free or obligatorily bound - the latter by combining with another morpheme to form a word (for example, *wis-* and *-dom* to form *wisdom*[5] - Bauer, 2007: 13). In most cases, words formed by combining obligatorily bound morphemes are neoclassical compounds, e.g. *cardiopulmonary* (from Greek *καρδία*, 'heart' and Latin *pulmo*, 'lung'). Classical languages can act as reservoirs of combining parts and are exploited by modern languages to form compounds, derivatives, and hybrid forms. Once these have been created they act as examples for further neologisms from their constituent parts (Dardano, 1994; Ross, 2004). Importantly, these combinations are not found in the language(s) of origin (Plag, 2018).

Morphosemantic analysis of INNs is possible because their meanings are highly compositional, in that their meanings are derived from the meanings of the constituent parts (e.g. Deléger et al., 2009), though not all of these parts are meaningful. INNs are not full neoclassical compounds to the degree of terms in anatomy and some other branches of medicine, because they cannot be parsed into elements that are directly derived from classical languages. Instead, multiple sources are drawn upon to create the names - a

[5] What kind of *dom* is this? Or is it a type of wis?

"random, fantasy prefix", with a STEM and possibly also a SUBSTEM. STEMS can be abbreviations (e.g. the SUBSTEM *-tu-* in situximab, for tumorous targeted tissue, or the STEM -kin in ilodekakin denoting interleukin-type substances) or acronyms (e.g. the STEM -mab in urtoxazumab denoting monoclonal antibodies).

In most cases, INNs are either partial neoclassical forms, drawing upon elements of chemical nomenclature (e.g. the STEM -fos, from Latin phosphorous, in clofenvinfos, denoting phosphorous derivatives), or they adapt a STEM that is based on a neoclassical element, as in the above -kin in ilodekakin, combining Latin *inter*, 'between' and Greek *leukos*, 'white', 'between white blood cells'). Thus, INNs can be seen as *quasi-neoclassical compounds*.

There is a huge pan-European lexicon caused by neoclassical word formation, and comparative studies have found many similarities between the morphological structures of neoclassical compounds in European languages of Indo-European origin (Booij, 2012; Iacobini, 2004; Magris & Others, 1992; Ross, 2004). INNs form an international nomenclature that is translated into six languages and is used as the basis for translation into many more languages. In line with the WHO naming guidelines (given earlier in this Chapter in Table 2.1) drawing upon neoclassical elements that are present in other languages facilitates international use and translation of INNs (Bryan et al. 2015).

2.2.4 INNs as meaningful units: semantic transparency

Semantic transparency in a complex word is formally defined as: "a measure of the degree to which the meaning of multimorphemic combination can be synchronically related to the meaning of its constituents and the typical way of combining the constituent meanings." (Schäfer, 2018). Research into semantic transparency in English and many other languages focusses on compounds (such as BLACKBIRD), and there is very little academic literature on semantic transparency of medication nomenclature (e.g. Bryan et al., 2015), or even on neoclassical compounds. This presents both a challenge and an opportunity. The challenge is that there is no experimental background or *a posteriori* knowledge on semantic transparency in INNs, with assumptions being tentatively inferred from research on general-language compounds and neoclassical compounds. Many general-language compounds contain free standing lexical items (e.g. BLUEBERRY, composed of BLUE and BERRY) whereas the constituents of INNs only exist in the names (Bauer, 1998). Findings from this analysis present the opportunity to contribute a new element of knowledge to this field of study.

One of many long-standing debates in linguistics is over the extent to which our brains carry out morphological processing of complex forms and the extent to which linguistic units are stored in the mental lexicon, as whole words vs. constituents vs. a combination of both (Bell, 2011; Bell, 2012; Biber et al. 1999; Frisson et al. 2008; Giegerich, 2004; Quirk et al. 1985; Scalise, 1986; Selkirk, 1982; Schäfer, 2018). For example, do we break down a compound such as BLACKBIRD into constituent morphemes BLACK and BIRD, and process the utterance as a BLACK type of BIRD? Or is BLACKBIRD stored as a complete word in the mental lexicon and not broken down into constituents? If language were to prioritise storage efficiency over computational efficiency, then multi-morphemic words would not be stored as composite wholes but in terms of their constituent morphemes. But storage efficiency of this kind would create a larger computational burden - multi-morphemic strings to be processed would need to be decomposed into constituent morphemes by some sort of morphological parsing procedure, and there would need to be linkage between these constituents (in some fashion). The alternative is the prioritisation of computational efficiency over storage, whereby multi-morphemic words are represented in their full forms in the mental lexicon, requiring more storage but less computation (Feldman et al., 2004; ten Hacken, 1994; Libben, 1998; Namer & Zweigenbaum, 2004). While the processing debate is not central to the thesis, it has implications for understanding how semantic transparency influences the mental lexicon.

In other words, how is the meaning of a complex word affected by the meanings of its constituents, and the ordering of those constituents ('synchronicity'), within the context of the typical way in which these constituents are combined?[6] In general language, transparency links forms to meaning, but in medication names it has particular significance as transparency points to specific medications and at times medical treatment. The key difference between general language compounds and medication names is that these word parts have been assigned a meaning within the nomenclature rather than in a system of word formation that has emerged naturally in a language (Ten Hacken & Panocová, 2015). The guidelines (WHO, 2017a) have been designed to create an INN that is both maximally transparent and sufficiently distinct, in a way artificially simulating the process of morphological decomposition by attempting to build INNs in a predictable way, using

[6] For example, BEDROOM is transparent: both BED and ROOM have predictable meanings, and they are ordered in a way that an English speaker can interpret the meaning as a type of ROOM, for a BED (although, they are not fully compositional or predictable and will vary by language, e.g. SCHLAFZIMMER, ‘sleep room’). On the other hand, HOGWASH is fully opaque, it is not a type of WASH nor a type of HOG and the meanings of neither constituents are predictable.

combining forms with a predictable meaning (Bryan, 2015; Bryan et al., 2015; ten Hacken & Panocová, 2015).

The order of compositional words parts of INNs is important for their recognition so it is important that a user can identify the 'head' of an INN as its semantic category. Neoclassical compounds are typically right-headed, though this is usually determined semantically, not syntactically - e.g. BIOGRAPHY is a type of *GRAPHY. Experimentally, a semantically transparent head can reduce time needed to process the compound's meaning, and compounds with opaque heads (e.g. JAILBIRD, HOGWASH, which are metaphorical) take longer to process (Libben et al., 2003). Thus, it is important that a user can first identify the head of a medication name, then deduce its meaning, and interpret its relationship to the other parts of the name. INNs as *quasi*-neoclassical compounds are also typically right-headed (found in Bryan et al. 2015), placing the STEM at the end of the name, as in montelukast with STEM -ast (denoting *antiasthmatic* medications) and SUBSTEM -lukast (denoting *leukotriene receptor antagonists*). This right-to-left display of taxonomy, from the end of the name to the beginning, is a predictable approach for the user as they can first classify the name under its superordinate (hyperonymic) concept, then under the hyponym, by reading to the left. Parsing the name from left to right would be futile in this case, owing to the random 'PREFIX' at the beginning of the name. The ways in which STEMS and SUBSTEMS are combined and ordered in INNs is presented as a typology in Section 2.4, expanded from earlier work (Bryan, 2015). Headedness is explored in this study on a larger sample of INNs than used in previous work (Bryan et al., 2015).

2.3 Terminology usage in this thesis

This section explains how certain linguistic terms are adapted for pharmacological use in the context of INNs and are given a special notation in the thesis to ensure clarity and consistency. Documentation on naming INNs published by the WHO uses certain standard linguistic terms in a non-standard way (WHO, 2017a; 2021b). To clearly distinguish this usage, these terms are in SMALL CAPS throughout the thesis. In particular, the terms STEM, SUBSTEM, and AFFIX (and types of AFFIX: PREFIX, INFIX, SUFFIX, and FREEFIX) have a specific pharmacological sense and a WHO-specific sense in this thesis and in the domain of INNs that is different to their linguistic sense. For example:

> *"INN for monoclonal antibodies (mAb) are composed by a random prefix, an infix, which indicates the target (molecule, cell and organ) class, and by the stem -mab as a suffix" (Stem Book, WHO, 2021b:p207)*

The linguistic stem is the base of a word, often serving as the core of its meaning (e.g., *walk* in 'walking'). It can usually be 'free', e.g. *walk.* The pharmacological STEM is a word part that conveys pharmacological meaning in a medication name. It functions as the semantic head of an INN and is 'bound', in the sense that it cannot appear alone (e.g. -ast has no pharmacological meaning outside of montelukast). In the mathematical linguistic sense, the pharmacological STEM it is acting more like a morpheme.

Linguistic affixes modify a word's meaning or class, such as 're-' in 'revisit' or '-ing' in 'walking.' Pharmacological AFFIXES indicate the position of the STEM in the INN, such as the PREFIX STEM cef- in ceftriaxone or the SUFFIX STEM - cept in aflibercept. This sense is extended here to FREEFIX STEM or simply FREEFIX, to describe a STEM that can appear anywhere in the name without a change in meaning (e.g. nab in cannabidiol).

WHO guidance describes a "random, fantasy prefix" intended to create a distinctive name. In practice, this is a misnomer, as the so-called PREFIX may appear in diverse positions within the name. For instance, in ceftriaxone, the supposed PREFIX (STEM) is actually a SUFFIX (STEM), and the random part is -triaxone. A revision of terminology in WHO guidelines would better reflect the variability in STEM positions in INNs.

2.4 The INN nomenclature as a taxonomy, and a typology

Medicine uses several large-scale systems of classification (Segura-Bedmar et al., 2008; L. G. Shapiro et al., 2005), such as:

- HUGO, a gene nomenclature
- MeSH, used to index published research on MedLine and PubMED
- UWDA, an anatomy nomenclature

The need for large systems of classification has arisen following the proliferation of research and information in all fields, but particularly in medicine, coupled with the need for professionals to stay abreast of current developments (Coletti & Bleich, 2001; Segura-Bedmar et al., 2008). An example of such an ontology is the UMLS, an automated piece of software that reads and assimilates information in published research (Segura-Bedmar et al., 2008), and bootstraps information within complementary systems, such as the UWDA. Some systems determine nomenclature (e.g. HUGO) and others assign semantic relations between concepts (e.g. in the UWDA, the oesophagus is *part-of* the foregut, *continuous-with*

the pharynx and stomach, and *adjacent-to* the trachea and thoracic aorta and thoracic vertebral column (Shapiro et al., 2005).

The terms *classification*, *taxonomy*, and *ontology* are often used interchangeably to refer to any system of categorisation, but in this thesis *ontology* denotes any system that categorises concepts, whereas a taxonomy should be seen as a methodology for categorisation. There are several key distinctions to be made. An ontology is “the concrete form of a conceptualisation of a community’s knowledge of a domain”, whereas a taxonomy does not necessarily include added knowledge beyond the necessary and sufficient criteria for categorisation (Stevens et al., 2000). Ontologies may be multidirectional and include multiple types of semantic relation, such as meronymy and metonymy, whereas a taxonomy is an upside-down tree structure and is based upon intrinsic properties of its members. Taxonomies are typically ‘tree-like’ hierarchies, employing hyponymy (Is-a, class membership) to express semantic relationships (Stevens et al., 2000). Within certain linguistic theories, the organisation of systems will inevitably depend upon our perception and conceptualisation of the world (Jackendoff, 1983) and therefore be inherently subjective. The plant and animal kingdoms described in biology constitute examples of taxonomies (Coletti & Bleich, 2001; Shapiro et al., 2005).

The WHO calls the INN nomenclature a 'classification', but it can be more specifically seen as a taxonomy, because it only uses hyponymic semantic relations. The WHO uses a unique taxonomy for classifying INNs, which does not align with the global taxonomic system for pharmaceutical substances, the Anatomical Therapeutic Chemical (ATC) index. Table 2.2 gives the example of selegiline. The INN taxonomy is at most four levels deep (although names currently only use two of these, so the taxonomy is relatively flat), and assigns alphanumeric codes to each level. The INN system is unique in the breadth of its usership, not restricted to use by healthcare professionals, but also used by patients, parents, and carers, and other professionals with varying degrees of expertise.

*Table 2.2 Comparing classifications of selegiline in INN and ATC systems**

selegiline - INN code C3120	***selegiline - ATC code N04BD01***
C000 Psychopharmacologics	N Nervous system
C300 Antidepressants	N04 Anti-parkinson drugs
	N04B Dopaminergic agents
C310 Monoamine oxidase inhibitors type B	N04BD Monoamine oxidase B inhibitors

* WHO 2021a:34, WHO 2021b

The pharmaceutical nomenclature and its peripheral systems of nomenclature (such as pathology, anatomy, nosology, etc.) are large and complex. A taxonomic conceptual system that is developed over time will inevitably contain some broken links and general inconsistencies, but these inconsistencies should not work to the detriment of the overall aims of the system. The structure of the INN taxonomy does not conform to the archetypal tree structure, and is at times messy and fractured.

An archetypal taxonomic STEM entails a clear tree hierarchy of concepts, based on hyperonyms (parent categories) and branching out to hyponyms (child categories). Although many names display their taxonomy in a predictable right-to-left ordering, e.g. montelukast (parent STEM -ast and child SUBSTEM -lukast), there is no single rule governing the position of STEMS and SUBSTEMS in names, and thus the display of taxonomy. The burden on users, then, is high - they need to understand the meaning of STEMS and layout of taxonomy, and to learn the many exceptions. Properly structured taxonomies can help to bring substantial order to a model, whereas improperly structured taxonomies can make models confusing and difficult to use. A robust taxonomy is also important for automated recognition systems, and this will increase the life-span and usage of the INN nomenclature (Segura-Bedmar et al. 2008; ONS 2024).

Previous research (Bryan et al., 2015, Bryan, 2015) revealed several problems with the INN taxonomy of STEMS and its implementation. We found wide variation in spelling of STEMS and SUBSTEMS, very orthographically similar but unrelated STEMS and SUBSTEMS, and an intricate and messy array of affixation, in which in some cases the position of the STEM motivates meaning, and in other cases has no effect on meaning, with no way for the user to distinguish between these without reading the entire 239-page WHO Stem Book (WHO, 2017a). SUFFIXES are the most common type of STEM, with SUBSTEMS acting as INFIXES (for example, montelukast in the -ast antiasthmatics taxon, in Table 2.3).

Table 2.3 The -ast antiasthmatics taxon

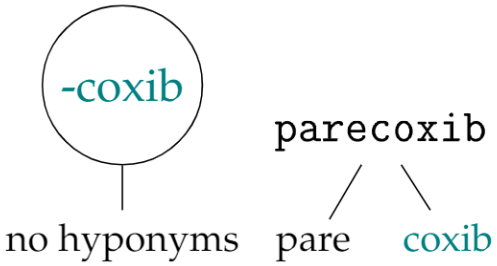
SUBSTEM (hyponym)	STEM (hyperonym)	Mode of action	Example INN	Therapeutic indication
(with no SUBSTEM)	-ast	Antiasthmatics or antiallergics, not acting primarily as antihistaminics	zaprinast	Selective PDE inhibitors, precursor to PDE-5 inhibitors such as sildenafil (Viagra)
-lukast		Leukotriene receptor antagonists	montelukast	Treatment for reversible bronchoconstriction
-milast		Phosphodiesterase type IV (PDE IV) inhibitors	cilomilast	Treatment for chronic obstructive airway disease and psoriasis
-trodast		Thromboxane A-2 receptor antagonists, antiasthmatics	seratrodast	Long-term management of asthma
-zolast		Leukotriene biosynthesis inhibitors	quazolast	Mediator release indicator

This right-to-left display of taxonomy, from the end of the name to the beginning, is a predictable approach for the user as they can first classify the name under its superordinate (hyperonymic) concept, then under the hyponym, by reading to the left. Parsing the name from left to right would be futile in this case, owing to the random PREFIX at the beginning of the name. In our previous research (Bryan et al., 2015), we explored formal properties of over 7,000 INNs, and semantic properties of a 1% random sample (n=71), and their associated STEM taxa. INNs do not employ a single type of taxonomy, but rather several. Previous work classified the taxonomy of INNs into four types, updated and described below (Bryan, 2015).

2.4.1 Single-level taxa

Many INNs are regularly formed, and employ a STEM representing a single taxon with no 'child' taxa, or hyponyms. Examples of these single-level taxa are arte-, -coxib, -formin, nab, and -pris-. These STEMS occur in all positions in the name: as INFIXES, FREEFIXES, PREFIXES, and SUFFIXES. The presence of single-level taxa illustrates the longevity of the INN nomenclature: the taxonomy has been designed to be future-proof from developments in pharmacology, by leaving room for categories that don't yet exist, similar to Dmitri Mendeleev's periodic table. Figure 2.2 gives an example of this taxon type, SUFFIX -coxib meaning 'selective [COX2] cyclo-oxygenase inhibitors'.

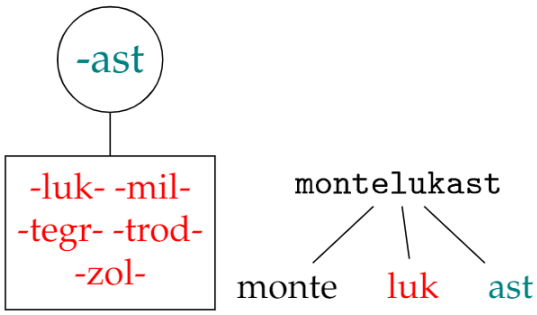
Figure 2.2 Type 1: Single-level taxon -coxib, for 'selective [COX2] cyclo-oxygenase inhibitors', and example parecoxib



2.4.2 Multi-level taxa[7]

Many STEM taxa have clear hyponymic relationships between 'parent' and 'child' concepts, and display taxonomy clearly in names from right to left. The STEM is the SUFFIX, and SUBSTEMS are distinguished from their co-hyponyms as INFIXES directly before the SUFFIX STEM. There is clear, linear, and predictable morphemic concatenation and the user can work right to left to decipher meaning. The taxon for “antiasthmatic, antiallergic substances not acting primarily as antihistaminics” has STEM -ast, and SUBSTEMS -lukast, -milast, -tegrast, -trodast, and -zolast. Montelukast is a substance in this group, and its meaning can be easily derived from the order of word parts: with SUFFIX STEM -ast it can be categorised as part of the antiasthmatic taxon, and INFIX -luk- as a leukotriene receptor antagonist. Figure 2.3 lays out the -ast taxon, complementing Table 2.3.

Figure 2.3 Type 2: Multi-level taxon -ast, for 'antiasthmatic, antiallergic substances not acting primarily as antihistaminics', with examples montelukast and roflumilast



[7] So-called ‘Regular simple’ type in Bryan 2015

2.4.3 Complex taxa[8]

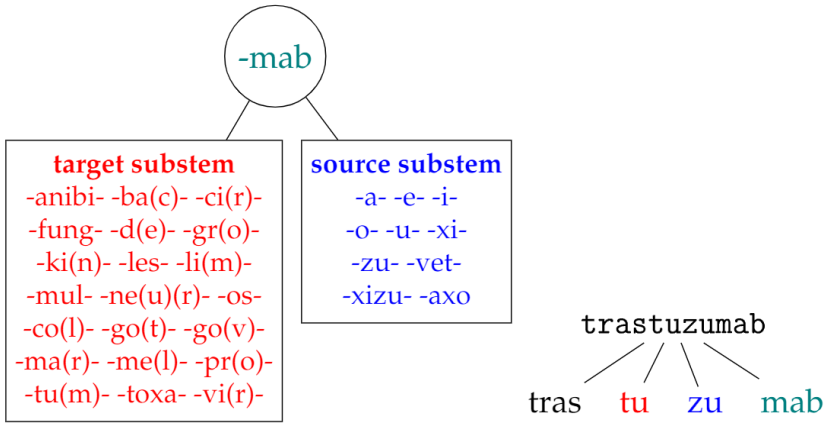
Monoclonal antibodies are a relatively new but rapidly expanding branch of pharmacology, with an INN taxon that is regular and well defined, but complex. These have clear hyponymic relationships between 'parent' and 'child' concepts, but are laid out over multiple levels or require complex morphemic concatenation, such as -mab in trastuzumab, an acronym SUFFIX STEM meaning monoclonal antibody. There is a regular correspondence between names in this group and their meaning, and meaning is predictable for the user. Up to April 2017, INNs for monoclonal antibodies comprised a random PREFIX (1), followed by two INFIX SUBSTEMS (2 and 3), and a SUFFIX STEM (4), referring in a specified order to:

1. random PREFIX;
2. SUBSTEM INFIX - target class or disease class;
3. SUBSTEM INFIX - the source class on which the immunoglobulin sequence is based; and
4. SUFFIX STEM -the hyperonym -mab.

Each of the three meaning-bearing elements must be defined in the name and they are concatenated in a certain order, but they combine freely with each other. This typology was first defined in a 2015 publication, and so was based on the taxonomy at that point. Figure 2.4 shows the taxon for monoclonal antibodies up to and including April 2017 (list 77), during which time 67 names for monoclonal antibodies were designated. With the exception of the first monoclonal antibody, muromonab, the naming system for monoclonal antibodies has changed twice since it was introduced - in April 2017 with the publication of list 78, the source SUBSTEM was discontinued, and in 2022, the target SUBSTEM was discontinued, and the -mab STEM was replaced with four distinct SUBSTEMS: -tug, -bart, -mig, and -ment (Balocco et al., 2022; World Health Organization, n.d.). Monoclonal antibodies in the full dataset therefore belong to both type 2 (lists 78 onwards) and type 3 (lists 1 to 77). From 2022 - not in this dataset - they belong to type 1, as they have no SUBSTEM.

[8] So-called ‘Regular complex’ in Bryan 2015

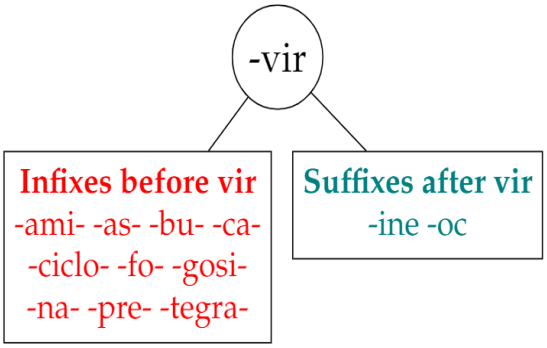
Figure 2.4 Type 3: Complex taxon -mab, for 'monoclonal antibody' and example trastuzumab; old system up to April 2017



2.4.4 Irregular taxa

In these taxa, the formation of STEMS and SUBSTEMS in a name is not predictable or in line with formation of names in regular taxa, such as the FREEFIX STEM vir (Figure 2.5) for antivirals, with SUBSTEM -viroc in maraviroc and -ciclovir in aciclovir. Unlike the very clear concatenation rules assigned to monoclonal antibodies, in which the distinguishing part of the hyponym is prefixed to the STEM of its hyperonym in the word, the formation of SUBSTEMS for hyponyms of a main STEM is not consistent in INNs, given the variation in types of AFFIX STEM used. The taxon for antivirals, with the STEM vir, will be used as an example. As a FREEFIX, the STEM vir can appear anywhere in the name, meaning that the user cannot rely on a right-to-left display of taxonomy to derive meaning. The STEM has defined SUBSTEMS -amivir, -cavir, -ciclovir, -fovir, -gosivir, -navir, -previr, -virine and -viroc, but names given as examples under the hyperonym also include ribavirin, enfuvirtide, and efavirenz - these names conform to the naming guidelines, but when vir is used as a FREEFIX it is difficult for users to parse their meaning.

Figure 2.5 Type 4: Irregular taxon vir, for antivirals



Examples: oseltamivir, bictegravir

Examples: etravirine, maraviroc

In an irregular taxon such as vir, there are two orders of morphemic concatenation: linear right-to-left as in aciclovir in Figure 2.6 or nonlinear, with the STEM vir appearing in the middle of the name, as in maraviroc in Figure 2.7. Since the right-to-left morphemic concatenation occurs in the other types of taxa, the meaning of aciclovir can easily be derived from the order of morphemic concatenation in the name. However, in maraviroc the distinguishing part of the hyponymic SUBSTEM -oc follows its hyperonym vir in the name, and thus it is not immediately obvious to which taxon the name belongs. One could mistake the name under the STEM *-oc (which doesn't exist). This can also be seen in other STEM taxa, such as the hyperonym prost with hyponym -prostil, and in any taxon for which the STEM is not a SUFFIX. This problem arises from the phenomena of FREEFIXES in INNs. Since STEMS such as vir and prost do not have a set position in the word, semantic transparency may be deemed low, since the user has no predictable way of recognising the main STEM.

Figure 2.6 Type 4: taxon vir, for 'antivirals' and example aciclovir

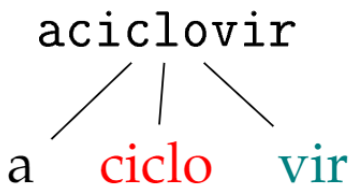
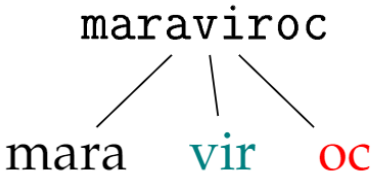


Figure 2.7 Type 4: taxon vir, for 'antivirals' and example maraviroc



2.5 Conclusion

This background chapter introduced the nomenclature under focus - International Nonproprietary Names (INNs) - and outlined the process of name formation, along with the naming guidelines stipulated by the World Health Organization (WHO). It contextualised INNs as linguistic items, introducing them as part of a nomenclature, quasi-neoclassical compounds, and as meaningful units with both formal and semantic features. The chapter also discussed terminology usage, highlighting how certain terms, such as STEM differ in meaning across the fields of linguistics and pharmacology. Finally, it situated INNs within a taxonomy and updated the typology of INN taxa, first proposed in Bryan (2015).

Chapter 3: Literature review

3.1 Introduction and background

This chapter critically examines the literature on LASA errors, expanding on themes introduced in prior publications, including a narrative review published in the *British Journal of Clinical Pharmacology* in 2020 (Bryan et al., 2020a). The chapter also draws on a systematic review of LASA error interventions, published in the same journal and based on a PROSPERO-registered protocol (Bryan et al., 2016, 2020b). These two reviews focus on drivers and solutions to the problem of LASA errors, offering foundational insights, but they are limited in scope in key areas such as linguistic underpinnings and identifying key gaps in the literature base that merit further attention. The current chapter adopts a narrative approach to complement previous work, allowing a broader exploration of contributions from multiple disciplines and theoretical perspectives. In synthesising insights from linguistics, pharmacology, and patient safety, persistent gaps in the field are identified. Furthermore the chapter critiques methodologies and assumptions behind current research, including prior work by the author. Search strategies are detailed in Appendix A2, with a flowchart summarising the search methodology in Figure 3.1.

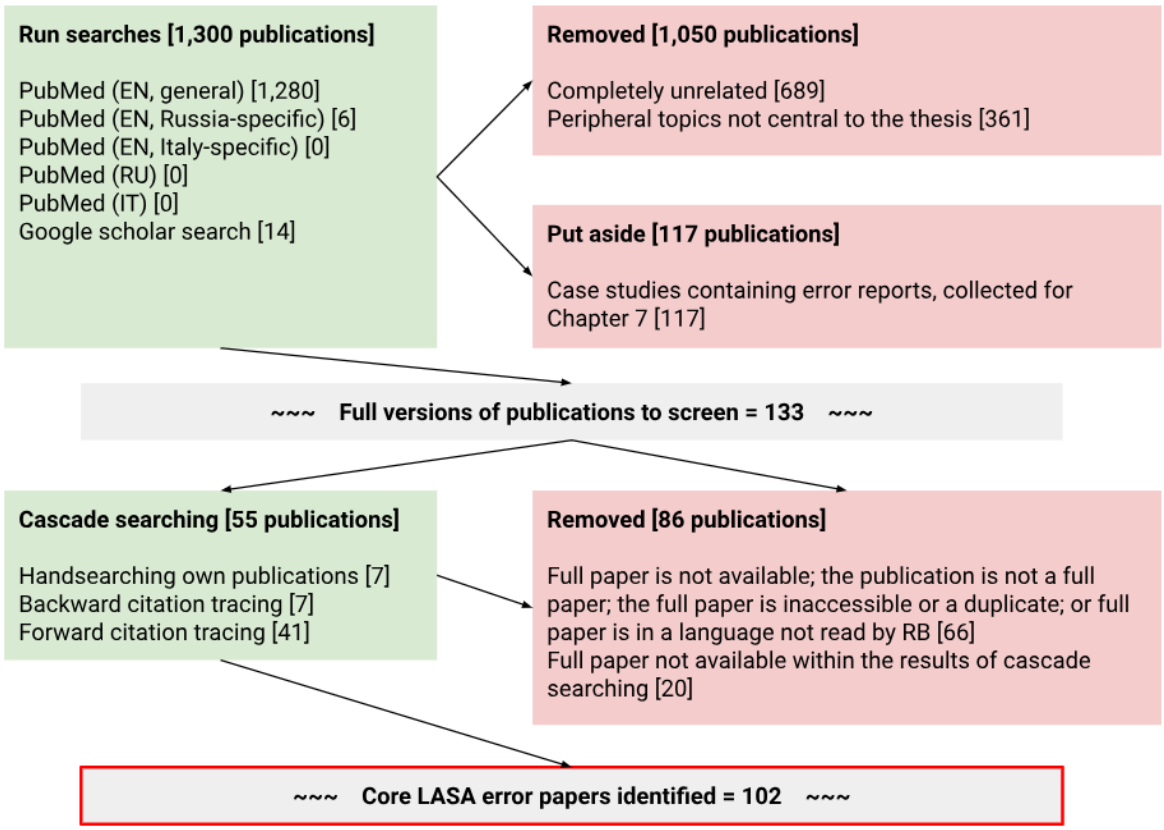
3.1.1 Search strategy

This narrative review uses a broad-spectrum search strategy (see Appendix section A2) designed to capture a wide range of literature on LASA errors. Unlike the systematic review (Appendix A2), this search excludes terms specific to interventions, allowing for a more comprehensive overview of the problem itself. While this broader approach successfully identified key studies central to the issue, it also revealed significant geographic and linguistic gaps, particularly in publications in languages other than English. For example, targeted searches for Russian- and Italian-language literature – using translated search terms and wild cards on PubMed – yielded no results for Russian and only six for Italian, none of which offered substantial insights into the problem of LASA errors in those respective countries. This points to a limitation in LASA error research, suggesting that critical perspectives from non-English-speaking regions may be underrepresented in the global literature.

Because there is such a broad range of terms used to describe LASA errors, a step-wise approach was adopted to selecting papers from the literature to explore further, using citation tracking to review forward and backward citations. After identifying papers on

PubMed that are central to the problem of LASA errors, the forward citations of those papers were then explored (later papers that cite the paper) and backward citations (papers in the past that it cites) and identified more relevant literature. A handsearch was also conducted of prior publications (Bryan, 2015; Bryan et al., 2015, 2020a) and overall thesis bibliography and added a few more. The process is illustrated in Figure 3.1.

Figure 3.1 Search methodology for Chapter 3



102 papers were identified for the thesis literature review, and these were read and annotated to varying degrees and various themes were explored as they emerged. This iterative process revealed new themes and gaps in the literature, yet it also highlighted some reliance in the field on a limited set of highly cited papers, which raises concerns about potential echo chambers in LASA error research. For example, of the 102 publications (all peer-reviewed journal articles), 17 are co-authored by Bruce Lambert and a further seven by Jeffrey Aronson, who is an external advisor to this study. Overall the upward trend in publications each year is promising and demonstrates growing multidisciplinary interest in the topic (Ciociano & Bagnasco, 2014). The growing body of literature also reveals gaps: for example, while Aronson's work provides a robust pharmacological perspective, it

underexplores psycholinguistic aspects; and Lambert's contributions to patient safety frameworks focus on systematic issues but lack focus on semantic (linguistic) aspects.

3.1.2 Structure of the literature review

This chapter delves into some of the themes found in the literature on LASA errors (in the search strategy described above), and highlights gaps in the literature, some of which this thesis seeks to address. These themes are presented and critically discussed in a structured way in this chapter, to provide a robust description of the research problem of LASA errors, as well as some ways that LASA errors are classified and the ongoing efforts to reduce LASA errors. Themes were broadly divided into four types: background information on LASA errors (definitions, prevalence, and classifications); theoretical aspects (human error theory, error causation, seriousness and harm); drivers of LASA errors (both linguistic and non-linguistic); and solutions to LASA errors (assessing similarity at the pre-approval stage, mitigation strategies in healthcare, reporting). These four types are critically appraised in this chapter, in Sections 3.2, 3.3, 3.4, and 3.5, respectively.

Sections 3.2 and 3.3, respectively, present the background to the research problem of LASA errors, and a critical overview of theoretical aspects of LASA errors that were identified using the search strategy. Section 3.2 seeks to answer the questions "what are they?" and "how often do they happen?". It synthesises the findings of publications that work to bring more information to light on LASA errors and – more broadly – medication errors, refining definitions, estimating incidence, and estimating prevalence of LASA errors based on a proportion of medication errors (e.g. Aronson and Ferner, 2009a; Bundy et al., 2009; Elliot et al., 2018; Emmerton & Rizk, 2012). It then goes on to place LASA errors within key error classification models (e.g. they are 'wrong drug' errors within the 'five rights' framework) and to address the lack of publications on LASA errors between generic names (e.g. Bundy et al., 2009).

Section 3.3 introduces and problematises key theories around human error and error causation, such as the 'person' and 'system' approaches and the hot cheese model, and contextualises LASA errors within these theories, looking to answer the questions "how do LASA errors happen" (Li & Thimbleby, 2014; Reason, 2000). A blame-shifting dynamic was frequently referenced and so this section discusses the burden of responsibility for LASA errors, asking the question "whose problem are they?" (Aronson, 2004; Cohen, 1995). The section then identifies three axes, or scales, of LASA error used in the literature – probability, opportunity, and seriousness – and debates how the shifting of each axis can dynamically

change the risk of LASA error (Lambert, Chang & Gupta, 2003). For example, high similarity between two names increases the probability of LASA error, but if the drug is rarely prescribed then opportunity is low. Finally, Section 3.3 considers the role of harm in LASA errors (e.g. Basco et al., 2016; Ferner & Aronson, 2006).

Sections 3.4 and 3.5, respectively, synthesise findings from publications on the *drivers of* and *solutions to* LASA errors. This expands work published in Bryan et al. (2020a). Section 3.4 presents the strengths and limitations of findings on four core types of linguistic features that drive the risk of error, that were identified using the search strategy: orthographic similarity; neighbourhood density; prescribing frequency; and shared strings (e.g. Lambert et al., 2001b, 2002, 2003; Shah et al., 2017). The section briefly presents non-linguistic drivers, such as similar packaging and users' familiarity with the name, but they are outside the scope of the study. Section 3.5 categorises solutions into a visual framework through the lens of the hot cheese error causation model, from upstream solutions that reduce the risk of LASA error before a medication is approved, to downstream solutions 'at the sharp end' in clinical settings, and learning and reporting. Section 3.6 summarises the findings of the literature review and lists the gaps identified that the thesis works to address.

3.2 Background to the problem of LASA errors

In this section the definitions of *medication* error and *LASA error* are refined, as a type of medication error. Findings are presented from studies on incidence, prevalence, and economic burden of medication errors and LASA errors, using short illustrative case studies. LASA errors are framed within popular error classification frameworks such as the 'five rights' and assess the preponderance of publications on errors between brand names in the literature.

3.2.1 Medication errors

Of all events reported to cause patient harm in the UK, medication errors are the most common. Between January and March 2018, they accounted for 10.7% of incidents (total 206,485) and 63 deaths (Jordan & Kyriacos, 2014; Ostini et al., 2012). Medication errors are particularly common in intensive care, paediatrics and neonatology, care of the elderly, anaesthetics, and obstetrics (Kaushal et al., 2001; Ostini et al., 2012). Some medication errors will result in overdose, adverse drug reactions, or under-treatment, and cause serious harm to patients (Aronson, 2009c; Moorman & Carr, 2008; Runciman et al., 2003). As more drugs enter the market, with greater variation in routes of administration, this problem is becoming increasingly complex (Nute, 2014).

In this thesis the following definition will be used to denote a medication error:

> *An unintended failure in the drug treatment process that leads to, or has the potential to lead to, harm to the patient.*

This definition is adapted from Aronson and Ferner (2009a), with the addition of 'unintended' and 'drug'. It is used by the European Medicines Agency (EMA, 2020), and was formulated according to sound principles of definition: it must not be circular; neither too wide or too narrow; use as simple language as possible; and be positive, saying what the definiendum is rather than what it is not. Instead of listing all the situations in which an error could occur (an extensional definition), the definition is intensional, whittled down to the necessary and sufficient conditions to qualify a concept. In this case, a *medication error*:

1. **Is a failure**, either to complete a planned action or in the action that is planned;
2. **It occurs at a stage in the treatment process;**
3. **It involves a drug;**
4. **It has the potential for harm to the patient.**

This definition excludes errors that occur during manufacturing and packaging, by isolating the error to the treatment process. The work of Aronson, Ferner, and Velo has been instrumental in the normative process of standardising terminology in medication safety, to ensure consistency in communication within the field. They have critiqued how definitions are often formulated using design by committee and handing down results as 'ex cathedra statements' with no explanation of thought processes behind them (Aronson & Ferner, 2005). In this thesis many of their definitions are used (such as *drug*, *excipient*, *substance*), and these are referenced in the glossary in A1.

Aronson and Ferner also address ambiguities in terminology, such as distinguishing between *prescribing* and *prescription*. *Prescribing* refers to the process of deciding which drug to prescribe, while *prescription* refers to the written or verbal order for that drug. Errors in *prescribing* may take the form of under- or over-prescribing, irrational or inappropriate prescribing; errors in *prescription* refer to mismatches between the intended and actual prescriptions, such as an incorrect patient name, incorrect drug name, incorrect dose, route, etc. (Velo & Minuz, 2009). The term *medication* is often ambiguous because it can refer both to the process of giving medication as well as the object undergoing the process, specifically

the drug. In turn, *medication error* may ambiguously refer to (a) an error in the process of giving a drug or (b) an error relating to the drug (Aronson, 2009a).

In this thesis, *drug* and *medicinal product* will be used interchangeably to denote the object, with the accepted definition in the glossary in Appendix section A1, adapted to include placebos and other forms of medical interventions, such as inducing anaesthesia. A drug may or may not comprise an active component (the drug); it also includes blood products, embryonic stem cells, attenuated viruses for immunisations, contrast media, etc.. Prescribable medications are referred to as components of therapeutic regimens. In this thesis terminology usage will be aligned to MeSH: ‘medicine’ refers to the practice and theory of medicine and ‘drug’ is normally restricted to adjectival use, e.g. ‘drug-induced’. MeSH terms without exception use ‘drug’ in relation to ‘adverse’, and so terms such as ‘adverse drug reaction’ (attributive use) will also continue to be used.

Lack of a universally recognised definition of a *medication error* complicates discussions on this issue and makes it difficult to quantify the prevalence of medication errors and the proportion of all medical errors that are medication errors. A systematic review by Lisby et al. (2010) found 26 different definitions of *medication error* across 45 studies from nine countries, with reported prevalence ranging from 4% to 75% of prescriptions, depending on the definition used. However, they found no association between definitions of error and prevalence. Multiple and conflicting operational definitions also complicate and weaken inter-study comparisons (James et al., 2009; Ross, 2004).

3.2.1.1 Prevalence and economic burden of medication errors

Medication errors cost lives, with prescribing errors estimated to kill more people every year than car crashes, workplace injuries, or airline crashes. While airplane travel results in only 0.43 fatalities per million miles, drug-related errors cause 1,250 fatalities per million patients (Galvin, 1998; Newcomer, 2000; Propp, 2000). Lambert argues that medication errors will always make up a high proportion of all errors, since drug therapy is the most common form of treatment (Lambert et al., 2002). On the other hand, the WHO framed many medication errors as ’preventable’ in their Global Patient Safety Challenge (GPSC; WHO, 2017b).

Following the launch of the GPSC in 2018, the UK government commissioned a rapid review and synthesis of the evidence on medication error prevalence and economic burden on the NHS (Elliott et al., 2018). They identified 36 studies with prevalence ranging from 0.2% to 90.6% of prescriptions. The considerable range in prevalence, from 0.2% to 90.6% of

prescriptions, can be explained by differences in the type of medication error, measurement methods, clinical settings, data sources, population groups, healthcare professional groups, and drug types. To elaborate, one study included in the rapid review looked at error rates of prescriptions written by pharmacists and found 10 errors in 509 (0.2%) prescriptions (Onatade et al., 2017). In contrast, another study included in the rapid review found that 96 of 106 (90.6%) patients in a care home – all of whom had Type 2 diabetes – were prescribed a 'potentially inappropriate medication' (PIM), which was classified as an error (Andreassen et al., 2016). The stark difference here is caused by methodological differences, in these cases between sample populations: the study by Onatade et al. (2017) looked at members from the general population; the study by Andreassen et al. (2016) looked at an acute setting of the elderly with a chronic disease.

Nevertheless, the rapid review authors found support for the prevalence rates in the broader literature. For example, rates were higher in elderly populations, in care home settings, and during transitional care, and the effect is confounded by co-morbidity such as dysphagia, kidney disease or dementia. The rapid review further also identified polypharmacy as a risk factor: one looked at 93 patients who had been the recipient of an error; each was prescribed an average of 10.9 medications (Friedman et al., 2007).

Section A4 in the Appendix summarises the rapid review's findings. It revealed that 72% of all errors posed little to no potential harm, while 26% and 2% had the potential to cause moderate and severe harm, respectively. Of that total 28% (66 million errors) that could cause harm, the majority (47 million) occurred in primary care. The authors noted that error rates in the UK were similar to those in comparable European health systems but also emphasised that many of the studies included were over 10 years old, so could be outdated due to changes in healthcare systems (Elliott et al., 2018). The rapid review also warned that they could not determine the proportion of errors that reach the patient and relayed the challenge of extrapolating harms from error data. They used adverse drug reactions (ADRs) as a proxy measure to calculate economic burden, but that will not have a direct one-to-one correspondence with harm (not all errors will lead to an ADR, and not all ADRs are caused by an error). They recommended introducing systems for routine data collection on medication errors and patient harm, and funding to support implementation of evidence-based interventions to reduce errors (Elliott et al., 2018).

In the UK, hospital and community pharmacies dispense 900 million prescription items each year, with an estimated 134,341 dispensing errors. However, over 85% of these errors are intercepted before reaching the patient (Irwin, Mearns, et al., 2011; James et al., 2009). A

study found that near misses (errors not reaching the patient) outnumber actual errors by a factor of six; 38 dispensing errors (0.08%) and 247 near misses (0.48%) of a total 51,357 items dispensed (Chua et al., 2003). When a medication error *does* result in patient harm, the consequences can include prolonged treatment, readmissions, and legal action against healthcare professionals (Irwin, Mearns, et al., 2011). The rapid review by Elliott et al. (2018) found that costs associated with medication errors varied widely, from €68 per intercepted asthma inhaler error (Zaidi et al., 2015) to €7 million for litigation claims related to anaesthetic errors (Cranshaw et al., 2009). The high variation in cost estimates raises concerns about the reliability of economic models measuring consequences of medication errors – factors such as underreporting, differences in healthcare settings, and indirect costs such as lost productivity and inactivity are not always adequately addressed (Elliot et al., 2018).

3.2.2 Look-Alike Sound-Alike (LASA) errors

Medication errors can occur when medications have similar-looking or similar-sounding names, and/or shared features of product packaging. These **wrong drug** errors are Look-Alike Sound-Alike (LASA) errors (Bryan et al., 2015, 2020a). They can occur during prescribing, dispensing, or administration of medicines, and can lead to administration of the wrong medication to the patient. They can result in overdosing, under-dosing, or inappropriate dosing (Aronson, 2009b). LASA errors make up a high proportion of all medication errors, with estimates ranging from 6% to 50%, and represent a significant threat to patient safety (Emmerton & Rizk, 2012). LASA errors can occur through the physical selection of the wrong medication (e.g. from a shelf) or digital selection from an electronic list (Emmerton & Rizk, 2012).

Reviews in the literature (Ciociano & Bagnasco, 2014; Emmerton & Rizk, 2012; Larmené-Beld et al., 2018) use the term *LASA error* to denote an error that is due to one or more of four possible factors:

1. look-alike name
2. sound-alike name
3. look-alike packaging or labelling
4. look-alike tablets, capsules, ampoules, or other forms of drug delivery

One study found that of 6,550 LASA errors: 55% (3,615) were due to similar written or spoken forms of names (Types 1 and 2); 27% (1,791) were due to similar packaging or labelling (Type 3); and 17% (1,144) were due to similar tablets or capsules (Type 4)

(Chanakit et al., 2013). Of a total 5,327 pairs of LASA names: 69% (3,695) were tablets or capsules; 18% (944) were injections; 5.8% (307) were liquids; 6.9% (367) were topical or external agents; and 0.26% (14) were chemotherapeutic agents (Chanakit et al., 2013). This thesis is primarily concerned with errors that are caused by Look-Alike Sound-Alike (LASA) names, nevertheless recognising that similarities in packaging and delivery can compound risk. Within names, drug name confusion is seen in four forms - only the first form is within the scope of this thesis:

1. Two different drugs with similar names
2. Two different drug formulations under the same brand name
3. One drug with two different brand names
4. Abbreviations of drug names, such as ISMN from isosorbide mononitrate (all Aronson, 2004)

It is prudent to add to this list errors in which two names are confused, but the result is mis-dosing. This may be due to: (a) confusion between the brand and generic names of the same medicine, such as oxycodone and Oxycontin (ISMP, 2019) resulting in double-dosing; or (b) confusion between two drugs, and wrongly inferring the dosage of the confused drug, such as morphine and hydromorphone (Tuohy & Paparella, 2005), resulting in mis/under/overdosing. Type (b) is within the scope of this study, when it concerns two generic names. There are examples of *double-whammy* confusions, where both the brand and generic names are similar, such as:

1. Valtrex/valaciclovir - Valcyte/valganciclovir
2. Viracept/nelfinavir - Viramune/nevirapine (Berman, 2004)

LASA errors occur because of shared linguistic properties (similarity) between two or more names, and lead to incorrect substitution of one drug for another. Similarity can be orthographic (written forms) or phonological (spoken forms), or both, and can be quantified using various similarity measures (Kovacic & Chambers, 2011; Lambert et al., 1999). Potential for a LASA error may be compounded by similarities between product packaging, dosage form, or strength (Emmerton & Rizk, 2012). Potential for similarity may also increase with an increased prescribing frequency of the medications involved and the number of medications with similar therapeutic indications (Lambert, Chang, & Gupta, 2003). Most LASA pairs are reciprocal (ISMP, 2019), i.e. each name in a pair has been mistaken for its counterpart. This reveals inherent pairwise similarity, rather than external environmental factors. Some examples are given in Table 3.1.

Table 3.1 Examples of LASA name errors

Medications involved	Details of error
Intended: mercaptamine; confused: mercaptopurine	In 2010 in the UK, the MHRA reported that an infant with nephropathic cystinosis was given mercaptopurine instead of mercaptamine. After taking the wrong medication for one month, the infant developed pancytopenia but made a full recovery after the error was noticed and rectified (MHPRA, 2010).
Intended: morphine; confused: hydromorphone	In 2005 in the USA, an emergency department nurse gave oral hydromorphone 10 mg (standard dose 1.3 mg 4 hourly) instead of oral morphine 10 mg (standard dose 5-10 mg 4 hourly) to an elderly man, a 3-5 fold overdose. She had been distracted by another patient almost falling out of bed. When she returned, she forgot to fill in the medicines reconciliation record for the cupboard. Also, hydromorphone was not usually stored in that cupboard, so she was not primed to be prepared for the confusion. The patient was discharged before the error was noticed and suffered a fatal respiratory arrest on his way home (Tuohy & Paparella, 2005).
Intended: vecuronium; confused: cisatracurium	In 2010 in the USA, cisatracurium was administered in error to a 1 week old neonate. The error was realised immediately and no changes in vital signs were observed (Sauberan et al., 2010).
Intended: fentanyl; confused: sufentanil	In 2000 in the USA, sufentanil instead of fentanyl was given to a 15-year old boy and a 45-year old man (who were admitted to the same ward on the same day). Both developed apnoea, cyanosis, flaccidity, and vital signs far below baseline (e.g. oximetry as low as 60%), but both made a full recovery. This was an approximately 6-fold overdose. For the 15 year old, 20 micrograms of IV fentanyl was prescribed, and supposedly administered – in actuality this was sufentanil in the same dose. For the 45 year old, 100 micrograms of IV fentanyl was prescribed, but 50 micrograms was supposedly administered – in actuality this was sufentanil (the same dose). Owing to similar names and packaging and the same manufacturer, they were put into the wrong medication drawer. The ampoules contained the same concentration (Chisholm & Klanduch, 2000).

There is no universally agreed definition of a LASA error, and Table 3.2 illustrates the range of published usages of the term. These examples show that (i) the property ’LASA’ is usually applied to medication name pairs, rather than to the error itself; and (ii) LASA names are viewed as the cause of the error. This is important - LASA name pairs are present in the nomenclature as a potential cause or source of error, regardless of whether an error has occurred (yet) (Emmerton & Rizk, 2012; Kondrak & Dorr, 2006; Lambert et al., 1999). In designing search strategies it is important to remember that LASA is a recently coined term, and much more of the literature uses alternative language, such as ’name confusion’ and ’similar drug names’, and even alternative acronyms such as SALA (Sound-Alike Look-Alike; Castle et al., 2005) or LASARA (Look-Alike Sound-Alike, Read-Alike; Zafar & Raza, 2020).

The search strategy in this review therefore accounted for inconsistency in terminology in this relatively recent field.

Table 3.2 Usage of terms LASA and Look-Alike Sound-Alike

Term in context	Source
Similarity between names as a cause or source of medication error	
"Similarity between drug names can cause errors in short-term memory as well as in visual and auditory perception."; "Patients sometimes receive the wrong drug because similarity in the spelling and/or pronunciation of drug names leads to errors in prescribing, dispensing and administration."	Lambert, Chang, & Gupta, 2003
"Some of the more common sources of medication errors are confusion between sound-alike medication names or look-alike medication names, and confusion due to similar appearances for medication packages, or similar labels for different medications."	Berman, 2004
"Medication errors commonly involve drug names that look or sound alike." "Similar-looking commercial labelling and packaging are common causes of medication error, often from the use of nearly identical packaging for two separate items."	Filik et al., 2006
"Many hundreds of drugs have names that either look or sound so much alike that doctors, nurses and pharmacists can get them confused, dispensing the wrong one in errors that can injure or even kill patients."	Kondrak & Dorr, 2006
"Around one in four medication errors has been attributed to orthographic (look-alike) and phonetic (sound-alike) similarity between drug names and/or look-alike or confusable packaging."	Emmerton & Rizk, 2012
"Considerable research on medication error has focussed on the provision of an incorrect drug to a patient caused by confusion between orthographically similar drug names or similar drug packaging."	Irwin et al., 2013
"The term LASA ("look-alike sound-alike") delineates a confusion of medication due to the similar labelling and packaging of different drugs, or similar labelling and packaging of the same drug containing different strengths."	Schnoor et al., 2015
"The potential for drug name similarity to cause medication errors is widely acknowledged by health care professionals."; "It is approximated that one-quarter of medication-related incidents voluntarily reported in the United States are caused by drug name confusion".	Emmerton et al., 2015
"Wrong-drug errors, thought to be caused primarily by drug names that look and/or sound alike, occur at a rate of about one error per thousand dispensed prescriptions in the outpatient setting and one per thousand orders in the inpatient setting."	Lambert, Schroeder, et al., 2016
"Medication names that look-alike [or] sound-alike (LASA) are the most common cause contributing to medication errors."	Wong, 2016
LASA errors as a hyponym of medication errors, confusion is the event	
"One of every four medication errors reported in the United States is a name-confusion error."	Lambert et al., 1999

"Confusions between drug names that look and sound alike account for between 15% and 25% of reported medication errors."	Lambert, Chang, & Gupta, 2003
"Confusions between drug names that look and sound alike are common, costly, harmful, and difficult to prevent."	Lambert, Bhaumik, et al., 2016
"Look-alike, sound-alike medication errors occur when the names of 2 [sic] drugs have orthographic similarity or phonetic similarity, forming a look-alike, sound-alike pair."	Basco et al., 2016

3.2.2.1 Prevalence of LASA errors

LASA errors make up a high proportion of all medication errors; estimates range from 6.2% (Shaw et al., 2013) to 15% (Bundy et al., 2009). A quarter of serious complications reported to the ISMP are due to confusion between similar names, and the ISMP estimates that only 1 to 2% of errors are reported (Kenagy & Stein, 2001). Wide variations in prevalence have been reported, in part due to poor reporting and inconsistent definitions (Berman, 2004)[9]. When stated as a proportion of overall prescriptions or orders, the prevalence of LASA errors is, of a magnitude, much smaller. According to available evidence, LASA errors occur somewhere in the range of up 3×10^{-5}% (Basco et al., 2010) to 2.2×10^{-3}% of all prescriptions (Joshi et al., 2007). Since around 1 billion prescription items are issued in the UK each year in community pharmacies, one could extrapolate that between 300 and 22,000 LASA errors occur each year in the UK, although a large proportion will either be near misses or be intercepted before reaching a patient. A summary of estimates is given in Table 3.3.

Table 3.3 Prevalence of LASA errors: a summary of prevalence noted or cited in the literature

	Paper	% (n/N)	Details
LASA errors as a proportion of medication errors or medical errors, or error events	Aschenbrenner et al., 1996	22% of errors reported; absolute numbers not available as the paper could not be retrieved; numerators and denominators not stated.	Multiple papers cite this report from the US Pharmacopoeia Medication Errors Reporting Program (MERP, date unknown but published in 1996); it reports that similarity in appearance or sound of medicinal product names played a key role in the commission of 22% of errors.

[9] Many papers state that LASA errors account for up to 25% of medication errors (Berman, 2004; Ciociano & Bagnasco, 2014; Emmerton et al., 2015; Lambert et al., 1999), although this seems to be a rather shaky upper ceiling, as the source cited for this figure is not an original study, but a list of potential LASA pairs (Davis et al., 1992, citing Lambert et al., 1999.

(including near misses)	Lesar et al., 1997	11.4% of medication errors due to similar names, both brand and generic (79/2103)	Evaluation of prescribing errors in a 631-bed teaching hospital and causes attributed; the figure of 11.4% includes 'wrong drug' errors due to similar names, wrong dosages, and wrong abbreviations, and so this is an upper estimate.
	Bundy et al., 2009	14.7% of medication errors (89/607)	This was an analysis of 607 error reports to US MEDMARX between 2003 and 2006; errors can be categorised under one or more different types; 105 of 607 reports pertained to a "wrong vaccine" error, of which 89 (14.7%) were errors within LASA groups (as defined by the authors, such as Td, Tdap, DTaP, and DT).
	Irwin, Ross, et al., 2011	19.2% (110/573 dispensing errors)	Retrospective analysis of 3,798 incident reports via DATIX from 23 Scottish hospitals over 5 years. Drug name similarity was most commonly cited as a contributory factor, but it is difficult to know the exact proportion of LASA errors that make up the 110 'incorrect drug' errors.
	Smith & Lesar, 2011	7% of medication near misses (143/2,044)	A systematic evaluation of 2,044 near miss prescribing events; 7% were due to sound-alike names.
	Shaw et al., 2013	6.2% of paediatric medication events (37/597)	Based on incident reports collected in 18 emergency departments in the USA; paediatric medication events accounted for 19% of 597 events; 6.2% (37) were due to LASA errors.
	Tsuji et al., 2015	17.9% of medication errors were due to wrong drug name (650 of 3,630 medication errors, of a total 758,310 inpatient prescriptions)	Retrospective study of medication errors in a hospital setting over an eight-year period.
LASA errors as a proportion of all prescriptions	Joshi et al., 2007	$2.2\% \times 10^{-3}$ of all prescriptions (244/113,346)	This was the proportion of errors attributed to brand name confusion in a general hospital in Delhi.
	Basco et al., 2010	$3\% \times 10^{-5}$ of all prescriptions (395/1,420,091 prescriptions)	A database rather than clinical outcomes; this US study identified 0.28 alerts to potential LASA errors per 1000 prescriptions; all the errors arose from 22 LASA name pairs prescribed over 6 months in South Carolina.
	Lambert et al., 2019	$1.2 \times 10^{-4}\%$ (61/488,481 orders)	Retrospective analysis of medication records to identify potential errors over an eight-year period. These were identified using an algorithm then reviewed by physicians.

Table 3.3 also illustrates the smaller proportion of overall medication events made up by LASA errors, from 0.3 to 22 in 1 million events (Basco et al. 2010, Joshi et al., 2007). Searches aiming to ascertain the prevalence of LASA errors indicate little evidence and a problematic reliance on secondary citations. The wide range in estimates reflects the range

of definitions of what constitutes a medication error, which hinders standardised error recording and cross-study comparisons.

Wrong drug errors are consistently cited as the most common type of dispensing error, and LASA names are consistently indicated as the cause of such errors – 19% in Chua et al. (2003), 36% in Ashcroft et al. (2005). In a literature review of dispensing errors, wrong drug errors were cited as the most common type of dispensing error (compared with wrong strength, form, quantity, or instructions), and similarity was listed as a contributory factor in 12 papers out of the 60 included studies (James et al., 2009).

3.2.3 Classification of error

There are many ways to classify medication errors, and indeed LASA errors, and it can be useful to look at the problem in more than one dimension. However, these classifications often operate in silos and fail to adequately capture the complexity of the error process. For example, the traditional '5 Rs' framework emphasises the importance of giving the right drug, to the right patient, in the right dosage, via the right route, at the right time (Bundy et al., 2009) and is frequently used to classify errors along the medication process from prescriber to patient. LASA error typically fall under "wrong drug" errors, due to visual or phonological similarities between drug names (such as hydromorphone–morphine in Tuohy & Paparella, 2005) or LASA packaging with or without a similar name (such as acebutolol and amiodarone packaged in identical drug blister packs in Pathak et al., 2004). The 5 Rs framework is useful but at times simplistic.

Focussing on only one dimension of the error process (e.g. prescribing or dispensing) risks neglecting interconnections between different stages, such as post-dispensing errors by the patient, or manufacturing errors before prescribing. In the same vein, Van den Bemt's classification lists errors in prescribing (the decision to select a drug to prescribe), prescription (the act of writing the prescription), transcription, dispensing, administration, and across-settings errors (van den Bemt & Egberts, 2007; Wong, 2016). This classification acknowledges the multi-step nature of medication errors but still compartmentalises them, failing to explore systemic failures that influence how these errors manifest. To return to Aronson's distinction between prescribing (the decision-making process) and prescription (the act of writing or typing the prescription), a LASA error can occur during prescription but not during prescribing, because a LASA error is unintentional (Velo & Minuz, 2009).

In an outpatient setting, the medication process is extended to include the patient's role. One root cause analysis of the medication errors in an outpatient setting identified five types of error: prescription error, the failure to provide a correct prescription (analogous to prescription and prescribing errors, and wrong drug error); delivery error, failure to deliver the prescribed medication to the patient (this will cover wrong route, wrong patient, wrong dosage, and wrong time); availability error, such as there not being enough of the drug for the specified time period; patient error, failure by the patient to use the drug correctly; and reporting error, failure of the patient to identify the correct drug, dosage, or frequency when reporting back to a physician (Friedman et al., 2007). A LASA error could take place in three of the five types here: in prescription, delivery, or by the patient.

Alternative frameworks have been proposed by Berman (2004) and Aronson (2009a), who classified errors by their origins or the cognitive processes involved. Berman focusses on systems failures, classifying the cause of errors as due to: communications deficits; knowledge deficits; performance deficits; or systems deficits (Berman, 2004). Aronson uses psychological theory to classify events, in order to explain them, rather than purely describe them. He classifies medication errors as knowledge-based errors; rule-based errors; action-based errors; or memory-based errors (Aronson, 2009a). In a knowledge-based error, one does not know or understand a particular piece of information, for example prescribing amoxicillin for a patient who is allergic. In a rule-based error, one misapplies or fails to apply a good rule, or applies a bad rule. In an action-based error, also called a slip, an action is performed that was not intended. In a memory-based error, also called a lapse, an action is performed due to forgetting a piece of information, such as administering penicillin when you have forgotten that the patient is allergic (Aronson, 2009a). Since LASA errors are often unintentional, they would accurately be classed as action-based errors within Aronson's framework, where a slip of the pen or a moment of distraction results in the wrong drug being administered (such as writing 25 mg instead of 2.5 mg). While these frameworks are more holistic in their approach, they are limited by their emphasis on the cognitive failure of an individual (usually a healthcare professional) rather than addressing organisational and structural causes of LASA errors. For example, the framing of a LASA error as 'action-based' neglects the wider issue of how insufficient training, busy and overwhelmed staff, and poor communication, may predispose healthcare professionals to such errors.

3.2.4 Types of drug names: generic and brand names

The nomenclature systems for medication names, whether brand or generic, can be a source of confusion and error. These names are not arbitrary; they are artificial terms

created by official bodies such as the WHO. They are morphologically complex and can be analysed into base components, or STEMS, which give information about the referent drug's chemical structure and pharmacological properties. The STEMS are determined by the naming body (the WHO), and they are formed from Latin and Greek roots or other methods. A drug has a generic name in each country in which it has marketing authorisation, and at least one brand name, often many. As Lambert et al. (2002) note, the sheer volume of names – often varying by country and manufacturer – presents a significant risk for confusion. This confusion is not merely theoretical. In a survey of anaesthetists, 80% (sample size of 86) reported meeting unfamiliar brand names regularly and only a third could correctly identify common drugs by their brand names (Bramley, 2009). The sheer number of brand names for some out-of-patent medications (such as ramipril, with 13 brand names) creates an overwhelming environment where clinicians and pharmacists may be more prone to errors, especially when under pressure or time constraints (Carney et al., 2011).

It is important to note the type of name involved in an error (brand or generic, or both) and the combination and direction of the error, so as to know what system of nomenclature is central to the error. Confusion may occur between:

1. Generic - generic names, such as penicillin - penicillamine
2. Brand - brand names, such as Prozac - Provera
3. Brand - generic names, such as Soriatane - sertraline
4. Generic - brand names, such as methadone - Metadate

On the face of it, there is not a tangible difference between a brand-generic name error and a generic-brand name error (numbers 3 and 4 in the list above), but in fact each warrants a separate subclass, even though the linguistic properties of each pair would be the same regardless of which name was intended and which was confused. If only brand names were confused with generic names, one might conclude that name use convention is a driver of error; for example, in the UK physicians are recommended to use generic names when prescribing and thus are more familiar with them, so they then reject a brand name in favour of a generic name. Because errors happen in all directions, it is clear that the properties inherent in the names themselves are larger error drivers.

Based on 968 LASA errors reported to the ISMP, brand names are more commonly confused than generic names: generic-generic name pairs account for 320 pairs (33.0%), brand-brand pairs account for 560 pairs (57.9%), and brand-generic or generic-brand pairs account for 88 pairs (9.1%, 43 and 45 respectively; ISMP, 2019), confirmed in Hoffman & Proulx (2003). This has led to a focus on brand name LASA errors in the literature (Basco et

al., 2010; Bulfone et al., 2012; Carney et al., 2011; Lambert et al., 2005; Shah et al., 2017), with confusion between generic names being overlooked. This focus, however, may obscure the less studied but equally important matter of confusion between generic drug names, an area that is critically underexplored. This study aims to address this gap by focussing the analysis on generic name confusion.

3.3 Theoretical aspects of LASA errors

This section introduces and problematises key theories in human error and error causation, such as Reason's 'person' and 'system' approaches, and the Swiss cheese model, and critically discuss who should shoulder the burden of responsibility for the prevention and mitigation of LASA errors. Three axes of error are presented that are frequently used in the literature to quantify the risk and the consequences for LASA error. Finally, the relationship with error and harm is discussed.

3.3.1 Error reduction: person and system approaches

Reviews of the problem of LASA errors commonly separate interventions into 'person' and 'system' approaches (Ciociano & Bagnasco, 2014; Emmerton & Rizk, 2012), a distinction first made by Reason (1990). The person approach commonly apportions blame by focussing on the role of the practitioner in error, implying negligence, carelessness, inattention, or incompetence and often leads to an oversimplified view of error causation (Ciociano & Bagnasco, 2014; Cutter & Jordan, 2013). The person approach focusses on the individual on the front line who commits the unsafe act and lays blame at their feet, citing forgetfulness, carelessness, negligence, recklessness, inattention, or moral weakness. Countermeasures under this approach aim to reduce variability in human behaviour, and include writing procedures, education programmes, CPD, promoting consistency with poster campaigns, disciplinary measures, litigation, and 'naming, blaming and shaming'. The underlying assumption of the person approach is that errors are avoidable.

However, the person approach is criticised for neglecting the complex, often chaotic environments where healthcare professionals operate, where external factors such as workload, stress, and interruptions play a significant role, especially in high intensity environments such as an emergency department (Tuohy & Paparella, 2005). Since a LASA error may occur in writing or speaking a name (language production) or in reading or hearing a name (language reception), it has been recommended that clinicians say and/or spell the name out loud, to ensure correct understanding and solidify the name in working memory (Emmerton & Rizk, 2012). Patient mortality due to medication error is linked to poor nurse-

patient ratios, job dissatisfaction, and burnout (Aitken et al., 2013), and errors are also attributed to long shifts, time pressures, work load, and interruptions (Rogers et al., 2004), and nurses' educational attainment (Aitken et al., 2013). While such interventions might reduce individual error, they do not address the root causes embedded in the system itself, limiting their overall potential to reduce error (Cutter & Jordan, 2013).

In contrast, the system approach, rooted in Reason's theory (1990), recognises that 'to err is human' and that the root causes of error lie in non-human factors present in the system (Cutter & Jordan, 2013; Reason, 2000). Errors are reframed as consequences of unseen systemic factors elsewhere in the chain. This approach attempts to reduce error by identifying and addressing **latent conditions** (Reason, 2000) in the system that prime the risk of error; this includes tracing an active failure (an error) upstream to its cause and identifying latent conditions. **Active failures** are unsafe acts committed by people at the 'sharp end' dealing directly with patients. Latent conditions are the resident pathogens in the system, caused by misguided decisions or mistakes made initially by designers, procedure writers, and management (Reason, 1990), and also human-related factors such as tiredness, overwork, (in)experience, and low job security (Aronson, 2009c; Cook & Woods, 1994). Latent conditions cause **error-provoking conditions** in the workplace that increase the probability of error. Latent conditions by their very nature may lie dormant for many years before combining with an active failure to form an **accident opportunity** - by identifying them before the error occurs, risk is managed proactively rather than reactively. Latent conditions, such as design faults, bad procedures, and inadequate alarm systems, also weaken the system more broadly. Countermeasures accept human fallibility and instead strive to change the conditions in which humans work. Thus, when an unsafe act occurs, this approach does not question 'who', but rather 'how' and 'why' the system defences and safeguards failed (Reason, 2000). A literature review of dispensing errors (James et al., 2009) applied this framework of error and found a "complex interweaving of active failures (slips, lapses, and mistakes), latent conditions, and error-producing conditions".

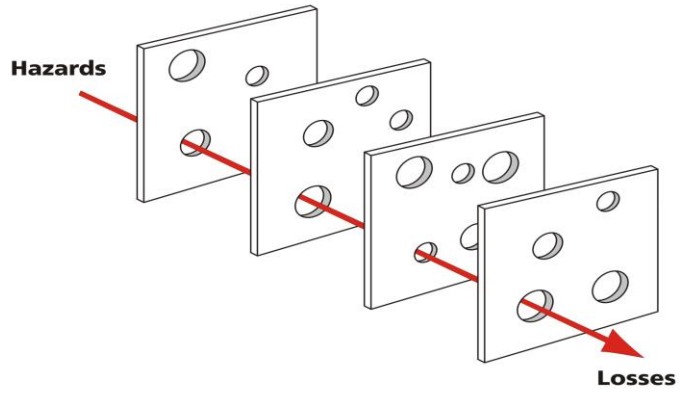
While more successful in reducing LASA errors (Emmerton & Rizk, 2012; Reason, 2000; Tuohy & Paparella, 2005), the system approach often encounters resistance due to its complexity and the challenge of shifting the responsibility from individuals to institutional structure. Organisations such as the ISMP encourage institutions to use "system-based strategies rather than depending on individuals alone to prevent errors" (Tuohy & Paparella, 2005). Despite the challenge, focussing on systemic factors – such as poor staffing ratios, high burnout, and underreporting due to fear of reprisal – has been shown to create a safer environment where errors are less likely to result in harm and encourages practitioners to

report errors and near misses, leading to greater understanding of error pathways and at-risk LASA name pairs (Aronson, 2009c; Emmerton & Rizk, 2012). Although a person approach may be more appealing for managers and in terms of litigation, continued adherence to this approach thwarts the development of safer healthcare institutions (Reason, 2000).

3.3.2 Models of error causation: Swiss cheese and hot cheese

The Swiss cheese model of accident causation, first developed by Reason (1990), offers a helpful, though simplistic, depiction of how multiple layers of defences in a system can fail in alignment, creating an opportunity for error. The multiple defensive layers are presented as slices of Swiss cheese with holes, as shown in Figure 3.2. The holes are **latent conditions** and are constantly changing shape and position. Latent conditions are the "potential contributing factors that lie dormant in the system and occur upstream at the more remote layers, called the blunt end" (Li & Thimbleby, 2014). Errors are active failures, and are viewed as the result of an accumulation of multiple failures in defences, whereby holes in the cheese unfortunately align to create a hazard trajectory and an error opportunity moving from left to right (Reason, 1990). For example, similarity between two names is a latent condition, and this causes the active failure of administering the wrong drug to a patient.

Figure 3.2 Swiss cheese model (SCM) - attribution David Mack / CC BY-SA



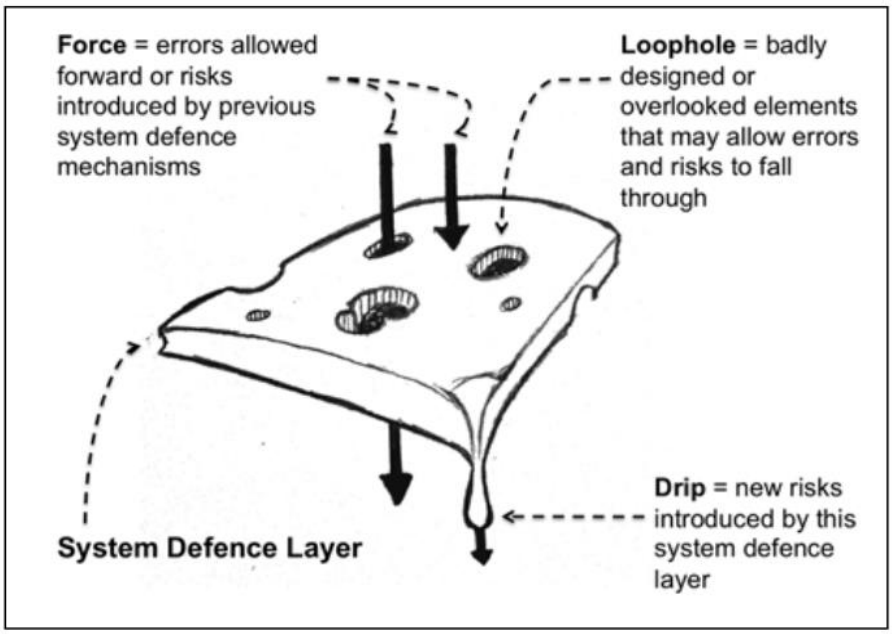
It has gained widespread use due to its visual simplicity and pedagogical clarity but it is also criticised for oversimplifying the complexity of error causation and neglecting interactions between defence layers (Bracco & Ivaldi, 2023; Dekker, 2002; Li & Thimbleby, 2014; Shappell & Wiegmann, 2000). The model fails to address the seriousness of consequences, which is critical in the context of LASA errors, where the implications of an error may vary

significantly depending on the drug involved, the patient's condition, and the timing of the error. Even its creator James Reason has questioned its usefulness (Reason et al., 2006). Its simplicity is a double-edged sword, meaning that it is easy to understand and use, but cannot capture the full complexity of error causation.

In response to these criticisms, the hot cheese model (HCM) presents a more nuanced approach (Li & Thimbleby, 2014). The hot cheese model, shown in Figure 3.3, arranges defence layers (slices of cheese) vertically and incorporates three key elements:

1. **force** - a combination of active failures and risks introduced from upstream defence layers, which will either be intercepted by a defence layer or seep through a loophole;
2. **drips** - new risks introduced by system defence layers due to design flaws, which may turn into force in the next layers downstream;
3. **loopholes** - latent conditions hidden in a layer that may allow *force* through.

Figure 3.3 Hot cheese model, reproducing Figure 2 from Li & Thimbleby (2014)



The hot cheese model allows for a more dynamic understanding of how latent conditions interact with active failures, particularly in complex systems such as giving medication. Since the latent condition of being 'look-alike, sound-alike' is found in multiple stages of the medication process, the HCM is useful to present various layers and interplay between them. For example, the model accounts for how an intervention (e.g. barcode medication

administration) can reduce errors in one layer but inadvertently introduce new risks elsewhere by prompting clinicians to adopt workarounds (Koppel et al., 2008; Ruuiainen et al., 2021). There are also some useful metaphors, such as the fondue pot acting as the incident reporting and learning sphere, collecting all the drips of hot cheese (c.f. literature on metaphor, e.g. Ritchie, 2013). However, the hot cheese model is not without its own limitations. The metaphoric 'drips' and 'loopholes' can oversimplify the nature of systemic risks, and it remains to be seen whether the model can fully capture the multifaceted interactions between defence layers in a high-stakes environment such as healthcare.

3.3.3 Who should bear the burden of responsibility?

The question of responsibility for LASA errors remains contentious, as healthcare professionals and pharmaceutical companies often shift blame to one another. The person approach of placing responsibility on practitioners, assumes a level of perfection that is unrealistic given the complexity of the medication-use process. Blaming individuals for errors also lowers reporting by creating fear of blame and reprisal (Berman, 2004). Pharmaceutical companies, on the other hand, are reluctant to implement comprehensive error reduction strategies due to concerns over liability and competition (Kenagy & Stein, 2001). This blame-shifting dynamic results in a fragmented approach to error prevention, where the responsibility for LASA errors is delegated to the frontline healthcare professional, despite the fact that the root causes typically lie in the design and manufacturing of drugs. This is especially true for LASA errors as a type of 'wrong-drug' error rather than 'wrong patient' or 'wrong route of administration'. More broadly, the problem of LASA name confusion should also be considered to be the responsibility of the manufacturers, regulators, and naming bodies (Aronson, 2004). Michael Cohen of the ISMP describes how:

> *Healthcare has a long history of tying serious medication errors primarily to the carelessness of practitioners while overlooking other major contributing factors, such as dangerous pharmaceutical packaging, non-existent warning systems, look-alike labelling, poorly conceived drug nomenclature, and faulty design of devices. …Unlike airline pilots and aerospace engineers, we in healthcare have all too often in responding to serious errors not designed appropriate systems to prevent such errors or inhibit their translation into fatal accidents (Cohen, 1995)*

As Cohen argues (1995), healthcare has failed to implement systems to prevent errors in the way that other industries, like aviation, have done. One study in aviation showed that 90% of errors were blameless, and policies such as giving pilots in the US immunity during accident investigations reinforces a no-blame culture (Marx, 1997; Reason, 1990). A holistic, collaborative approach is needed, one that involves all stakeholders – from drug manufacturers to regulators to healthcare professionals – working together to prevent LASA errors. A healthy reporting culture is essential and helps experts to understand the landscape of error and identify recurrent error traps.

However, there is little focus in the error literature on these 'upstream' stakeholders such as manufacturers and regulators, and in designating drug names there are competing goals: marketing for pharmaceutical companies; safety for regulators such as the FDA; the desire of the WHO and those who produce pharmacopoeias for memorable, clear, and usable names; and trademark protection for patent offices (Lambert et al., 1999). Although naming bodies, such as the WHO and the British Pharmacopoeia Commission (BPC), have a central role in mitigating the entry of LASA names into nomenclature, previous research has uncovered vague and unquantifiable WHO guidelines that are applied very inconsistently (Bryan et al., 2015). However, that 2015 study looked at the semantic features of a very small sample of INNs (1% of the nomenclature) so findings are tentative.

As a latent condition present throughout the system, the responsibility for LASA names cannot be borne by a single party, and a holistic approach is needed. A multifactorial approach involving all aspects of the medication-use process, from initial naming of INNs through to education and involving the patient (Ostini et al., 2012). The medication process is too long and complex for a single intervention to be effective at every stage, and instead requires a multimodal response (Lambert et al., 2022). Overall political commitment is also essential to overcome structural roadblocks (Ostini et al., 2012). When regulators, manufacturers, and naming bodies pass the buck, this increases the presence of latent conditions within the system and moves them 'downstream' to the sharp end, closer to the patient and within the realm of HCPs, who are more likely to be blamed for a resulting error.

3.3.4 Three axes of error: probability, opportunity, seriousness

The prevalence of LASA errors does not equate to the actual risk they pose to patients, because many errors are intercepted before reaching a patient. Instead, they are picked up downstream by pharmacists and nurses – they still happen, but they do not reach patients. This distinction between prevalence and risk is critical for assessing LASA errors, as it

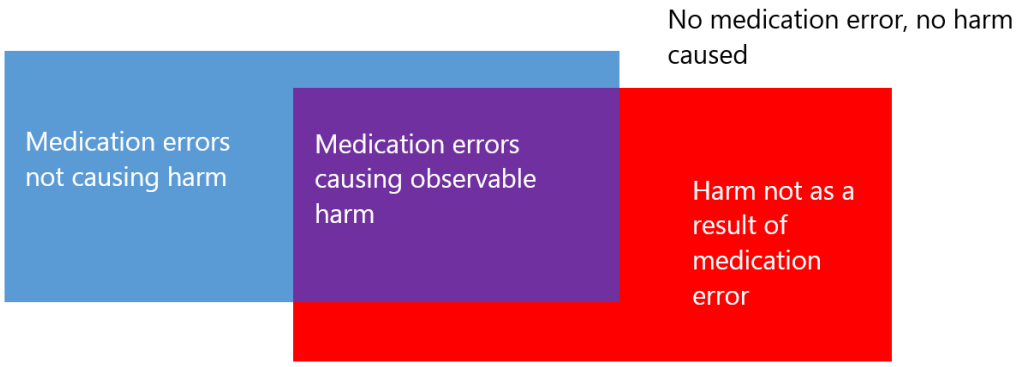
highlights the complexity of error dynamics within healthcare systems. Prevalence is the quantification of those errors that went undetected, reached a patient, and were reported or somehow identified retrospectively. For example, of the 650 wrong drug errors identified in one study, 73 were not picked up by pharmacists, of which only 9 were not picked up by nurses and reached patients (Tsuji et al., 2015). Many more errors occur at some point in the medication process.

The concept of risk can be accurately assessed along critical axes of error: **probability**, **opportunity**, and **seriousness** (developed from Lambert, Chang, & Gupta, 2003). This framework helps to evaluate the likelihood of an error occurring, the conditions that enable errors to happen, and the seriousness of the consequences if an error does reach a patient. These form a useful lens through which to assess risk and can fit within Reason's human error theory. Cross-referencing multiple axes of error is found in several studies, such as Emmerton et al. (2015), who used a matrix of seriousness (of outcome) and probability (in the form of a BI-SIM similarity score) to identify 35 of 250 pairs that pose an extreme risk of harm if they reach the patient; and a collaboration between the Institute for Safe Medication Practices (ISMP) and the Joint Commission on Accreditation of Healthcare Organizations (JCAHO) to create lists of LASA name pairs ranked by both their probability of confusion and seriousness if confused (Tuohy & Paparella, 2005). Clinical risk must be dynamically and continually assessed, because a drug prescribed more frequently has a lower probability of being confused, but owing to greater usage the actual incidence number will probably be high (Lambert, Chang, & Gupta, 2003). Errors that did not cause patient harm are less likely to be reported, which will skew the body of evidence and make it difficult to properly assess risk.

The probability of a LASA error occurring is influenced by factors such as the similarity between drug names and the frequency with which a drug is prescribed. High-frequency drugs present more opportunities for errors, even if their similarity to other drugs is low. Opportunity, then, is high when a drug is frequently prescribed, increasing the number of error-provoking conditions (Reason, 2000). An error may be extremely probable if two names are similar, but if it involves a rarely prescribed drug, it is unlikely to occur. The seriousness axis addresses the potential harm caused by an error, which depends on variables such as the drug involved, its dosage, and the patient's condition. However, the seriousness of an error is not always straightforward, as not all errors lead to observable harm in every instance and not all harm is the result of error (see Figure 3.4). Aronson argues that harm additionally implies a lack of benefit to the patient, which would include a patient not receiving a prescribed drug (Jeff Aronson's conference presentation, Sicily 2016).

Basco defines three types: harm from receiving the mistaken drug, such as intended methylphenidate but mistakenly given methadone; harm from not receiving the intended drug, such as intended furosemide but mistakenly given fosinopril; and harm from both, such as did not receive intended albuterol, and mistakenly given labetalol (Basco et al., 2016). Potential harm is key in the above definition of medication error by Ferner and Aronson (2006).

Figure 3.4 Relationship between error and harm: a Venn diagram



Any study of LASA errors should not overlook errors that do not reach patients. In a survey of healthcare professionals, respondents were asked to categorise errors as having reached the patient or not, as well as the intensity and seriousness of the potential or actual harm to the patient (Chanakit et al., 2013). The majority of errors that occurred did not reach the patient. Only 0.82% (26/3,161) resulted in any sort of harm to the patient, and none of this harm was life threatening or fatal. Other studies support this finding, such as Cina et al. (2006) in which 423 of 5,075 errors (8.33%) have the potential to reach patients, or Ashcroft et al. (2005), in which 280 of 330 incidents (84.85%) were classified as near misses. Ferner argues that the whole treatment process should be considered rather than outcomes alone, and the definition by Aronson and Ferner includes near misses or errors that do not lead to patient harm (Aronson, 2009a).

In any case, extra safeguards must be put in place if the seriousness of an error is high, as the consequences are more serious. Considering the seriousness of an error is also important in policymaking, because Tall Man lettering (described in detail in Section 2.5.2) has been shown to work better when used sparingly; it should be preserved for LASA pairs that would cause extreme harm. These axes provide a more nuanced understanding of risk,

but it also underscores the complexity of implementing effective interventions. Effective policy-making needs to prioritise safeguarding measures for high-risk drugs while addressing the root causes of error further upstream in the medication-use process.

3.4 Extrinsic and linguistic features of LASA name pairs

This section weighs up the strengths and limitations of findings on linguistic features that drive the risk of LASA error, identified during the literature search. The occurrence of LASA errors between medication names is influenced by various latent conditions, such as orthographic and phonological similarity, neighbourhood density, and prescribing frequency (Lambert, Chang, & Gupta, 2003; Lambert et al., 2001b). These features can interact and amplify the risk of errors (Reason, 2000) and are discussed in turn in subsections 3.4.1 to 3.4.4. Intrinsic linguistic features such as formal and semantic features of drug names are the focus of this study, and extrinsic factors such as familiarity with the name[10] and similar packaging, though they can increase the likelihood of confusion, are excluded (Her & Toh, 2019; Galanter et al 2014; Mcdougall et al., 2016). Prescribing frequency is treated as a linguistic feature here, as a proxy for word frequency (as in Lambert, Chang, & Gupta, 2003) and is discussed in section 3.4.3 below.

3.4.1 Orthographic and phonological similarity measures

Orthographic (written) and phonological (spoken) similarities between pairs of names have been shown to be accurate predictors of LASA errors (Abolhassani et al., 2017; Lambert, 1997; Lambert et al., 2001b, 2002). In psycholinguistic research, it has been demonstrated that similarity between words adversely affects recognition, recall, and perception (Anderson et al., 1998; Grainger & Jacobs, 1996; Lambert et al., 2001b, 2002; Luce & Pisoni, 1998; Wallace et al., 1995, 1998). Techniques such as edit distance (how many changes are required to turn one string into another) and n-gram matching (how many smaller strings, usually two-letter bigrams or three-letter trigrams, are shared) are commonly used to measure orthographic similarity, though these methods have limitations (Kondrak & Dorr, 2006; Lambert et al. 1999)[11]. These methods are easily transferable to measures of phonological similarity, by converting orthographic strings into phonological strings, using algorithms such as ARPAbet or ALINE (Jurafsky & Martin, 2000; Kondrak & Dorr, 2006; Lambert et al., 2010).

[10] Frequent misperceptions also occur when unfamiliar drug names are involved, suggesting that better familiarity with medical nomenclature could mitigate these risks (Lambert et al., 2010).
[11] There are also lesser-used measures, such as the longest common subsequence ratio (LCSR, calculated by dividing the number of characters in the longest common subsequence by the length of the longer string, (Kondrak & Dorr, 2006).

The Dice coefficient, for example, is used to quantify shared n-grams, such a three-letter trigrams, between two names, producing a similarity score between 0 and 1. For example, the trigram similarity between *mercaptamine* and *mercaptopurine* is calculated as 0.55[12], with six shared trigrams *mer, erc, rca, cap, apt,* and *ine*. While useful, these measures fail to capture all forms of similarity, and there are cases in which two different words would be deemed identical (such as nexan and xanex, with four shared bigrams; Kondrak & Dorr, 2006) – this can be counteracted by adding characters to the beginning or end of words, such as 'trigram-2b' measure, which adds two spaces to the beginning of the string and splits strings into three-letter sequences (Lambert et al., 1999). Even the best available models leave room for error, and recent advancements suggest combining orthographic and phonological similarity measures to improve accuracy and presenting similarity as a calculation of the mean (Lambert et al., 2019). A combined measure is useful because edit distance and shared ngrams pick up different types of similarity; one study found that only 17% of similar pairs were identified by both types of measure (Kovacic & Chambers, 2011).

Orthographic and phonological similarity measures form only part of the picture, but they are a working solution and are accessible, objective, and predictable (Lambert, 1997). However, even when measuring similarity of new proposed names, some will be missed. In the measure evaluation study by Lambert (1999), the multivariate model they tested had a sensitivity of 93.7% and a specificity of 95.9%. This translates to a false negative rate of 6.3% and a false positive rate of 4.1%. So when screening a new name against a database of existing names, the best model (at that time) would have missed roughly 1 in every 16 truly confusing names, and misidentified 1 in 24 as confusing (Lambert et al., 1999). Quantitative similarity testing is a key tool in the error reduction toolkit, but there are other confounding variables. Similarity is multidimensional - names may be similar orthographically, phonologically, and in handwriting or print. Beyond orthographic and phonological factors, other variables, such as handwriting or print style, and even language-specific influences (e.g., a user's first language), complicate the straightforward application of similarity measures in drug name safety (Anto et al., 2011; Lambert et al., 2005).

3.4.2 Neighbourhood density

Orthographic and phonological similarity measures have been the mainstay solution for LASA errors since Bruce Lambert's work in the late nineties, but more recently the concept

[12] 2 * 6 (shared trigrams) / 10 (trigrams in mercaptamine) + 12 (trigrams in mercaptopurine) = 12 / 22 = 0.55

of **neighbourhood** has emerged as an important driver of error probability. Neighbourhood density – the number of other names that are similar to a name – is a crucial factor influencing LASA error probability. Names in dense neighbourhoods are more likely to be confused due to their orthographic or similarity to other drug names (Irwin et al. 2011). While traditional similarity measures focus on pairwise relationships, neighbourhood density acknowledges that the risk of confusion is not just a function of two names but their relationship within the broader nomenclature (Lambert et al. 2005). The assumption of similarity being symmetrical and reciprocal is problematic (Lambert et al., 2005; Tversky, 1977). It appears reciprocal, because errors occur between pairs of names and are reported, but in lists such as that held by the ISMP, many names appear multiple times and have been confused with multiple other names, such as clonazepam with alprazolam, clobazam, clozapine, clonidine, and lorazepam (ISMP, 2019). This broader understanding of similarity as a network or neighbourhood, rather than a binary pair, represents an important shift in thinking about LASA errors.

In most definitions a **lexical neighbour** is a name that differs in only one letter from another name; thus a neighbourhood is a group of names that all differ by only one letter to a name; **neighbourhood density** is the number of names (Andrews, 1997; Perea & Rosa, 2000). In psycholinguistics the difference in a letter would typically be in the same position (i.e. *bin*, *bat*, and *pan* are all neighbours of *ban*), but due to this being too restrictive for drug names (compared with three-letter words in general language), Lambert and his research team extend this definition to all drug names with an edit distance of three letter, regardless of position (Lambert, Chang, & Gupta, 2003). A low neighbourhood density increases perceptual identification accuracy, and the converse is also true: names with high neighbourhood density (lots of similar names) are more likely to be misperceived than those with relatively few neighbours (Irwin, Ross, et al., 2011; Lambert, Chang, & Gupta, 2003). In visual perception tests, the presence of lexical neighbours reduces accuracy of drug name recognition, and having more neighbours amplifies this effect – multiple drugs with similar names should not be stored together (Irwin, Mearns, et al., 2011). Pharmacists are constantly tasked with searching a shelf or cupboard for a name with a dense similarity neighbourhood and multiple potential distractors, and it is crucial that names are not viewed in isolation or within pairs, but as part of a vast and multifaceted nomenclature system (Irwin, Mearns, et al., 2011). Neighbourhood density was found to be a competing factor with usability in Bryan et al., (2015), where names that are informative and easy to use are more likely to share a STEM and thus have more orthographic neighbours. The study also revealed that nearly half of INNs apply pharmacological STEMS in an inconsistent way and many names are distinguishable from other INNs by fewer than five letters, e.g. siplizumab,

ruplizumab, and teplizumab (so-called *neighbours*, Bryan et al., 2015). The 2015 study measured only density as a neighbourhood feature, and did not explore neighbourhood frequency.

3.4.3 Prescribing frequency and neighbourhood frequency

The frequency with which a drug is prescribed plays a role in its recognition and subsequent likelihood of confusion. More frequently prescribed drugs tend to be more familiar, thus less likely to be confused with others (Lambert, Chang, & Gupta, 2003). However, errors can arise when low-frequency drugs share high similarity with high-frequency drugs – in these cases, the less familiar drug may be misperceived due to its similarity to a more common one (Lambert, Chang, & Gupta, 2003; Lambert et al., 2010). Thus, an error is more likely to occur when a low-frequency drug is prescribed; on the other hand, the opportunity for this error is lower, since the drug is prescribed less frequently, so this risk is most salient when a drug has a significant level of similarity to other names, and a high neighbourhood density.

Prescribing/dispensing frequency is not ostensibly a linguistic feature; it can be used as a proxy measure for familiarity with a name (Lambert et al., 2005) and appears to assume the same effects as lexical frequency, i.e. that more frequently used words are more easily recognised, and thus frequently used drugs are less likely to be misperceived than rarely used ones (Lambert, Chang, & Gupta, 2003). Frequently prescribed/dispensed and familiar names are more accurately recalled than rarely prescribed and unfamiliar names, even when accounting for the effect of similarity – this further reinforces the effect of frequency, in that familiar names can be retrieved more easily from memory (Lambert, Chang, & Lin, 2003).

The effect of neighbourhood frequency on a name is mediated by the prescribing frequency of the name itself, regardless of the size of the neighbourhood. When a name is prescribed more frequently than the other names in its neighbourhood, then it stands out among its neighbours and is less prone to a recognition error. On the converse, when a name has a neighbourhood of high frequency names, it is drowned out and more susceptible to be confused. However, neighbourhood frequency alone does not appear to directly affect error rates; rather, it interacts with the individual drug's prescribing frequency to shape its susceptibility to confusion (Lambert, Chang, & Gupta, 2003).

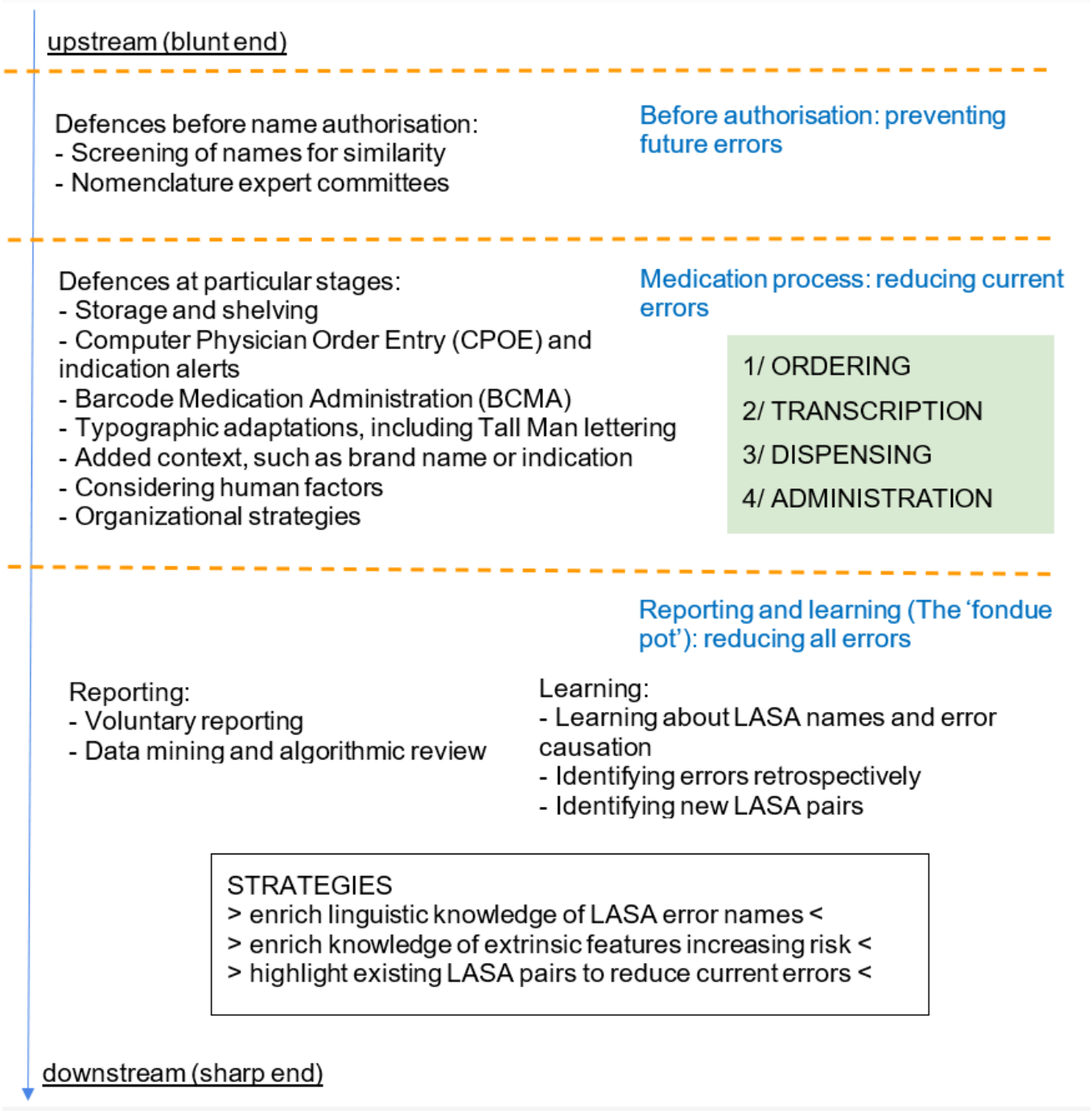
3.4.4 Shared beginning and end strings

Research suggests that drug names sharing the same initial and terminal strings are particularly prone to LASA error. For example, antimicrobial names beginning with ‘cef’ or ‘ceph’ have been frequently confused with one another (Thompson, 2008). In an evaluation of errors between pairs of brand names reported to ISMP, 68% of pairs analysed shared the same initial letter (most commonly <p>, <a>, <l>, and <z>), and 52% shared at least a trigram, typically at the beginning of the name. 91% percent of pairs had the same or a similar number of letters (Shah et al., 2017). These findings reinforce the practice of ‘weighting’ the beginnings of names in orthographic and phonological similarity measures, such as trigram-2b (Lambert et al., 1999).

3.5 Framework of the field of LASA errors

In this section a visual framework of solutions to LASA errors is presented, drawing on the hot cheese model (Li & Thimbleby, 2014) to outline a system of defences that range from ‘upstream’ drug naming processes to ‘downstream’ reporting and learning systems. The framework, presented in Figure 3.5, divides the field of LASA areas into three key areas: preventing new LASA names from being designated (before authorisation), reducing incidence of LASA errors during the medication process (after authorisation), and reporting and learning from past errors (constant and ongoing). Each of these areas is examined critically, with attention to their strengths and weaknesses.

Figure 3.5 Conceptual framework for Chapter 3



3.5.1 Before authorisation: preventing future errors

Preventing LASA errors from the outset by screening new drug names before approval is a proactive measure. It is useful to distinguish between (a) preventing errors by screening new names before approval; and (b) mitigating the risk of error between existing names (Filik et al., 2004). Scholars in the field tend to focus on one of these, such as Bruce Lambert on the former and Ruth Filik on the latter. It is clearly better to identify LASA name pairs before they are approved, as this will prevent errors involving those names (Lambert et al., 2001a; Lambert, Chang, & Gupta, 2003; Lambert et al., 2001a). The number of computations is much higher when scoring similarities between existing names than between a handful of proposed names (p) and existing names (n), as demonstrated by the number of scores for proposed names at the pre-approval stage in Equation 3.1, versus scores between all existing names at the post-approval stage in Equation 3.2. It is clearly preferable to identify similarity before a name is approved.

Equation 3.1 Number of similarity computations for 10 proposed names against 8,000 existing names

$$pn + \frac{p^2 - p}{2} = (10 \cdot 8{,}000) + \frac{10^2 - 10}{2} => \ 80{,}000 \text{ computations}$$

Equation 3.2 Number of similarity computations between 8,000 existing names

$$\frac{n^2 - n}{2} = example \frac{8{,}000^2 - 8{,}000}{2} = 31.9 \text{ million computations}$$

The number of LASA INNs coming into existence is reduced by the WHO INN committee via two mechanisms: measuring similarity between proposed names and existing names; and expert judgement of the INN Expert Group, who meet twice-annually to discuss and approve new INN requests (WHO, 2017a). There are a number of similarity measures, with varying degrees of sensitivity and specificity. Technical aspects of similarity measures are discussed in more detail earlier in this chapter. The WHO INN committee does not explicitly state the use of particular similarity scores in the approval process, but there are occasional references to ’search tools’ and ’comparisons’ with existing names and trademarks (SoINN, 2020a). In a 2016 consultation report, the POCA (Phonetic and Orthographic Computer Analysis) scores between two proposed names are given, indicating that the INN committee is using the tool (WHO INN committee, 2016). The POCA tool was created by the FDA and launched in 2002, and is an algorithm measuring the phonological and orthographic similarities of a proposed name (generic or brand) against multiple datasets of both brand

and generic names (Lambert et al., 2005; US Food and Drug Administration, 2007). The initiative has been hailed as a paragon of a successful top-down global policy, and industry manufacturers and regulators are encouraged to use it when proposing new brand names (Lizano-Díez et al., 2020).

On average, 96% of INN requests are approved by the WHO (SoINN, 2020b); some examples of INNs rejected because of similarity are given in Table 3.4. In contrast, another study showed that the FDA rejects around a third of proposed names because of the risk of confusion (Berman, 2004). Although tools such as POCA should reduce LASA errors involving new names, they cannot be used retroactively, and no INN has ever been amended after being 'recommended' (published) - unlike USANs, such as amrinone , which was changed to inamrinone after multiple confusion errors (Berman, 2004). There are also various psychological tests for evaluating the potential confusability of drug names before they are approved, and lab-based tests have been shown to be generalisable to real-world scenarios (Schroeder et al., 2017) – it is unclear whether such tests are being used by nomenclature approval committees.

Table 3.4 Rejected INNs

Stage of rejection	**INN requested**	**Too similar to**
Initial request	lunakimab	namilumab
Initial request	musontuzumab	sontuzumab
Initial request	adonikimab	etokimab
Objection to pINN[13]	deforolimus	everolimus
Objection to pINN	vanalimab	ianalumab
Objection to pINN	saridegib	sonidegib
Objection to pINN	votrisiran	Votrient ™

However, the WHO's reliance on tools such as POCA and on expert judgement is opaque and not universally trusted for predictive accuracy. The 'black box' nature of similarity scoring and the lack of transparency in decision making based on similarity limit its

[13] A name that passes the initial request stage and is reviewed by the INN committee is assigned the status of *proposed INN*, or *pINN*. When it is approved, it becomes a *recommended INN*, or *rINN*.

effectiveness. The extent to which the INN committee uses POCA is not known, but email communication with a member of the committee suggests that it was used in the past but it no longer used (communication with Dr Jeffrey Aronson, 2024). The accuracy of the similarity score and its ability to predict error is also heavily reliant on the exact method of the algorithm - since the underlying code is closed source, this is unknown. Furthermore, the threshold for acceptable similarity is unknown, so manufacturers are unable to self-assess their suggested INNs before submitting the request, which currently costs $12,000 per name. Overall assessment is a little vague and unquantifiable, as reflected in findings from a previous paper on INN naming guidelines (Bryan et al., 2015). On the other hand, the decision process is transparent - proposed INNs are discussed individually and in detail at each consultation meeting and written up in a report (Bryan et al., 2015; WHO, 2020c). Expert judgement seems to be the preferred method for the INN committee to assess similarity - INN documentation states:

> *The INN expert group assesses whether a new INN candidadate [sic] is too similar to an existing one. This is a subjective judgement depending on each INN expert's background in languages, the geographic and cultural context, as well as one's affinity or aversity to risk (SoINN, 2020a): slide15).*

Similarity scores and expert judgement before a name is approved can go some way to preventing more pairs of LASA names from entering the nomenclature, but there are thousands of existing names, some designated almost 70 years ago. The WHO themselves acknowledge that search tools are becoming more sophisticated and are thus able to identify more potential LASA names (SoINN, 2020a). Moreover, the psychological and cognitive challenges associated with LASA name confusion – particularly how similarity between drug names affects human memory and decision-making – are often underestimated. Although the pre-approval screening process aims to catch potential LASA pairs, the complexity of the clinical problem means that additional strategies are necessary at later stages in the medication process and these are discussed in the next section.

3.5.2 Medication process: reducing current errors

Once a medication has been approved and is in use, the focus shifts to reducing the likelihood of LASA errors in the medication process. At this stage, technological and organisational interventions become critical, and several strategies for reducing potential errors involving existing names have been proposed and will be discussed in this section. As illustrated out in Figure 3.5, some defences are in place at particular stages, such as

indication alerts in prescribing and dispensing, storage in dispensing, and barcodes in administration. Other defences do not target a specific stage, such as typographic changes to labelling, and consideration of human factors in design. The medication process is divided into five stages, as defined by Lisby et al. (2005), with an additional stage of *Self-Care* introduced, acknowledging the increasing role of patients and caregivers in medication errors (Akici et al., 2004; Brass & Weintraub, 2003; Lambert et al., 2001b). Table 3.5 outlines the stages of the medication process and their success criteria.

Table 3.5 Medication process, adapted from Table 1 in Lisby et al. (2010)

Stage	**Success criterion**
Ordering	Unambiguous prescription
Transcription	An identical copy of prescription in medical record
Dispensing	Dispensed medication is concordant with prescribed drug in nurse medication chart
Administration	The right medication to the right patient in the right way and at the right time
Discharge summaries	Eligible prescriptions in medical record are identical to prescriptions in discharge summaries
Self-care	The right medication taken by the right patient in the right way and at the right time

3.5.2.1 Indication alerts and Computerised physician order entry (CPOE)

Computerised physician order entry (CPOE) systems, a type of e-prescribing combined with clinical decision support (CDS), are integral in reducing medication errors, including LASA errors, by providing real-time alerts when potential LASA drug pairs are selected. Advanced forms typically match patient-specific data with a computerised knowledge base to generate patient-specific recommendations (Sim et al., 2001). The data from linked electronic health records can be tapped and used to retrospectively identify errors that have occurred, and add them to ‘watch lists’ in the software (such as Rash-Foanio et al., 2017). CPOE and CDS reduce medication errors and dosing errors, particularly when compared with no strategy (Bates et al., 1998, 1999; Brown et al., 2017; Garg et al., 2005; Kaushal et al., 2006; Nuckols et al., 2014; Roumeliotis et al., 2019). Some software can identify a mismatch between the drug selected and the patient’s electronic health record. For example, an alert may read:

> *This drug is typically used for hyperthyroidism. No such problem appears on the problem list of this patient –> [Cancel / Ignore / Add diagnosis]. (Galanter et al., 2014)*

Computerised alerts are used in various forms to varying degrees and are an effective defence layer against LASA errors in prescribing, working in real time (Emmerton & Rizk, 2012; Galanter et al., 2014). They can furthermore contribute to lists of problem names, jog attention, and inform about specific properties of LASA pairs, such as names that share the initial three letters, which are more likely to be confused. This interacts with the learning process, by identifying as-yet unknown LASA pairs of names and potential errors but stops them before they occur. Matching a prescription with the indication adds context and a further differentiating factor that will mitigate against a LASA error taking place (more about added context in the section below). In a survey of clinicians, 86% (230/269) had noticed an alert for drug name confusion, and 9% (20) of those said the alert had prevented them from dispensing the wrong drug (Campmans et al., 2018).

However, as the hot cheese model highlights, introducing a defence layer can create new openings for error (so-called *drip*); in the case of CPOE, users may choose to simply override computerised warnings and there is significant associated alert fatigue, so the intervention must be used strategically (Backman et al., 2017; Campmans et al., 2018). Human factors are an important consideration in the design of such systems, as inflexible systems lead users to create their own workarounds, which in turn may be error-provoking – however, one study showed that less than half of all CPOE system vendors employ usability staff (Ratwani et al., 2015).

3.5.2.2 Storage and shelving

It is recommended that LASA drugs be stored physically apart from one another to reduce the risk of selection, or picking, errors. In terms of psychology, selection of a drug encompasses two distinct processes: choosing the correct item and rejecting distractors (Irwin, Mearns, et al., 2011). Strategically organised shelves and storage areas can separate items with orthographically similar names, and this reduces the number of distractors in the user's visual field. Strategic storage will inevitably be an improvement on alphabetical ordering, since names that share initial two- or three-letter strings are more likely to be confused (Lambert et al., 2001a; Shah et al., 2017; Thompson, 2008). Amy Irwin has found that multiple similarly-named targets in the visual field reduces selection accuracy and increases reaction times (Irwin et al., 2013); another study showed that 45% of wrongly

selected drugs had names that were in close proximity to the prescribed drug (Anto et al., 2011).

3.5.2.3 Barcoding

Some studies have estimated that a third of all errors take place during administration, and are therefore more likely to reach the patient and to cause harm (Bates et al., 1995). Barcode medication administration (BCMA) technology has been developed to reduce medication errors at the administration stage, by confirming at the bedside the five rights of medication: right drug, at the right dose, at the right time, with the right route, to the right patient. Patients are assigned a barcode wristband on admission, and this is scanned before any medication is administered. BCMA can also be integrated into dispensing and is used to correlate physical selection of the drug with selection on the screen. This is used to varying degrees and is especially popular in Australia because of financial incentives (Emmerton & Rizk, 2012; Tuohy & Paparella, 2005). A key disadvantage to barcoding is that it risks reproducing errors that were made before dispensing, as these are less likely to be spotted if additional manual checks are not performed (Bryan et al., 2020a).

3.5.2.4 Typography, especially Tall Man lettering

Printed and digital names can be presented in various ways to maximise their readability and distinctiveness, by changing font colour, weight, kerning, and capitalisation. Tall Man lettering is a popular way of changing the typography of drug names; it has been endorsed by the US Food and Drug Administration since 2001 (US Food and Drug Administration, 2020), and is used in many hospital pharmacies in the UK (personal communication, ABMU pharmacist). Tall Man lettering uses selective capitalisation of letters in LASA name pairs, to highlight characters that distinguish them from each other, for example, DOBUTamine and DOPamine, or hydrALAzine and hydrOXYzine (Emmerton et al., 2015; Filik et al., 2006; Gabriele, 2006). Tall Man lettering has the potential to reduce LASA errors in written/typed, but not spoken, communications, and it differentiates look-alike packaging. Moreover, it is relatively easy to implement, both on physical packaging and electronically. Given that similarity between the desired selection (target) and other items (distractors) in a visual field increases search difficulty (Duncan & Humphreys, 1989; Farmer & Taylor, 1980; Irwin et al., 2013), it is beneficial to enhance certain properties of the target that distinguish it from distractors, so-called *feature-based processing* (Maunsell and Treue, 2006), such as changes in typography. Examples include bolding, italicising, colour contrast.

A systematic review of the efficacy of Tall Man lettering found that Tall Man is a marginally effective intervention to reduce LASA error, with a number of caveats (Bryan et al. 2020b). It

presented a Tall Man placebo effect, whereby users derive more benefit from the intervention when they are aware of its purpose, and found a ceiling of efficacy, beyond which in certain high-risk situations the risk of confusion cannot be mitigated by typography alone (Bryan et al., 2020b), but it is worth noting that a meta-analysis was not possible due to heterogeneity of the included studies, and fewer than ten studies were identified for inclusion.

3.5.2.5 Adding further context

Adding context to a name confirms the correct name and further distinguishes between similar names. Many studies suggest using the drug indication alongside the generic name in a prescription, and this has been shown to be effective; for example, participants in Seoane-Vazquez et al. (2017) were able to differentiate 26 of 33 LASA name pairs when they were accompanied by the FDA-approved indication. The drug indication is the semantic link between the patient's condition or diagnosis and the medication being prescribed, so this is helpful in both manual and computerised prescribing, to jog the clinician's attention and confirm the choice of drug. Most pairs of drugs with LASA names do not share an indication (59%, [165/268] according to Cheng et al., 2018), and therefore associating names with indications should go some way to differentiating a LASA pair; in cases where they do show some overlap, further context will be necessary (Cheng et al., 2018; Pairman et al., 2024; Seoane-Vazquez et al., 2017). It may also be useful to use the brand name in conjunction with the generic name, for added context and differentiation (Bramley, 2009).

3.5.2.6 Considering human factors

Considering human factors – ensuring that users are adequately trained, have clear guidelines, and work in environments conducive to error mitigation – is critical and can create systems that are more resilient to error and can absorb damage (Mokdad et al., 2021; Reason, 2000). Elements range from simple, such as allowing for illegible handwriting in written prescriptions to complex, such as centring the user in CPOE system design (Aronson, 2004; Brodell et al., 1997; Ratwani et al., 2015). Error-provoking conditions also include busy pharmacies, inexperienced staff, and understaffing (Allinson et al., 2005; Irwin, Mearns, et al., 2011; Irwin, Ross, et al., 2011). Efforts to reduce the risk of current errors are focussed on the medication process, from prescribing to administration, and involve people at every stage. Technological solutions are important, but human factors cannot be overlooked in the design of these solutions.

There is a lot of literature on reducing interruptions and distractions for healthcare professionals, so that they can focus on the medication task at hand. A LASA error can

occur in writing or speaking a name (language production) or in reading or hearing (language reception), so it is recommended that healthcare professionals say and/or spell the name out loud (before typing or handwriting or administering it), to ensure correct understanding and to solidify the name in the working memory (Emmerton & Rizk, 2012). A suite of processes to reduce distractions and interruptions during prescribing, prescription, dispensing, and administration have been proposed, such as "Do not disturb" tabards to be worn during drug rounds, and "no interruption" zones, with varying degrees of success (Alteren et al., 2018; Raban & Westbrook, 2014).

Humans are subject to bias, in particular confirmation bias - when a person looks at something and sees what you expect to see, rather than what is there. Just as when reading they generally only read the first and last letters and skim over the middle of the word, a pair of medicine names that share the first and last letters may easily be confused (Nott, 2001; Schulmeister, 2006). Confirmation bias is cited as a major contributor to LASA errors (e.g. in Tuohy & Paparella, 2005), and can be mitigated by having two people individually carry out the same task (Berman, 2004).

3.5.2.7 Organisational strategies

Reason emphasises the importance of studying successful organisations, rather than those experiencing failure. This includes recognising that human variability can be a benefit, insomuch as people can adapt their behaviour to changing situations and form a safeguard – this is ’dynamic safety’. These organisations can ’shapeshift’ in certain situations, pushing power towards experts and management during a crisis, then back to a ’normal’ conventional hierarchy when the crisis has passed. These organisations generalise failures, rather than localising them, and train staff to recognise error and recover from it (Reason, 2000).

It may be beneficial for organisations to adapt LASA error reduction strategies to each department, as each area of medicine uses a different set of core drugs, and so the LASA names to watch out for will vary. Since users experience alert fatigue, and interventions such as Tall Man are more effective when used sparingly, tailoring LASA watch lists by department or specialty will maximise impact (Backman et al., 2017; Campmans et al., 2018). There are some useful papers presenting LASA name pairs in oncology, which provide a template for identifying LASA watchlists in other specialties (Basco et al., 2016; Kovacic & Chambers, 2011; Schulmeister, 2006).

3.5.3 Reporting and learning, the 'fondue pot'

The final stage in the framework involves reporting and learning from errors. Incident reporting systems, such as the Yellow Card Scheme in the UK and ISMP in North America, play a critical role in collecting data on LASA errors. However, underreporting remains a pervasive issue, limiting the capacity to fully understand and address the scope of LASA errors (Aronson, 2009c; Evans et al., 2006; Li & Thimbleby, 2014; Shiima et al., 2022). Errors can also be found retrospectively through data mining or reviewing medical claims reports and electronic health records using tailored search algorithms and statistical text classifiers (Wong, 2016). While retrospective data mining and analysis can uncover trends in LASA errors, they are only useful to the extent that the underlying errors are documented in the first place. The 'fondue pot' metaphor in the **hot cheese** model illustrates the importance of collecting and consolidating errors for knowledge transfer and organisational learning (Emmerton & Rizk, 2012; Shaw et al., 2013). The fondue pot collects the 'drips' – the errors and near misses. By integrating real-world data into error reduction strategies, it is hoped that policymakers and regulators can refine their approaches and prevent future incidents.

There are three overarching goals of the 'fondue pot' for LASA errors. First, to enrich knowledge of the linguistic features of LASA names, in order to inform the pre-approval stage, and prevent future errors. Secondly, to enrich knowledge of the extrinsic features of LASA names, in order to inform risk reduction strategies and prevent current potential errors. The third goal is to identify and highlight LASA pairs of names existing in the nomenclature to prevent current errors. The first goal is the focus of this thesis: to enrich the body of knowledge about the linguistic features of LASA names.

3.6 Summary and Conclusion

The literature on LASA (Look-Alike, Sound-Alike) errors, first identified in 1973 (Teplitsky), has seen substantial growth, particularly since the late 1990s, with a significant contribution from Bruce Lambert. This chapter reviewed the literature through a comprehensive search, providing an overview of LASA errors in terms of their definition, classification, prevalence, and economic impact. Despite the considerable body of work, much of the research remains concentrated on orthographic and phonological similarities, with less attention paid to the semantic features. The scope of the thesis was narrowed to the study of LASA errors between different drugs with similar generic names, distinguishing this from errors linked to packaging or visual similarity (although these may also be a cause in some error incidents).

Three fundamental axes were introduced for the classification of LASA errors: probability, opportunity, and seriousness. This conceptual structure was used to identify the linguistic features most likely to contribute to errors. This framework highlights a multi-layered approach involving prevention, mitigation, and learning. While pre-approval screening and technological interventions offer promising solutions, they are not without their limitations. The complexity of LASA errors, influenced by human cognition, organisational factors, and technological design, demands ongoing attention and improvement across all stages of the medication process. And the need for robust reporting and feedback systems cannot be overstated. A more transparent, holistic, and evidence-based approach to error prevention, incorporating both technological and human factors, is essential to reducing the prevalence and impact of LASA errors.

Drawing on Reason's person-system approach to error causation, here systemic rather than individual blame is emphasised, underscoring the latent conditions within the pharmaceutical system that facilitate errors (Reason, 2000). This perspective aligns with the growing consensus in the literature that systemic factors – such as the structural design of drug names and their display – play a pivotal role in the likelihood of LASA errors. This system approach frames humans as fallible and error as caused by latent conditions in the system that must be identified and safeguarded against (Emmerton & Rizk, 2012; Reason, 2000; Tuohy & Paparella, 2005).

The chapter identified a gap in the existing research, namely the exclusive focus on formal linguistic features (orthography and phonological) with little exploration of the semantic aspects of drug nomenclature. The lack of studies investigating the semantic properties of drug names is notable, especially considering that the beneficial effects of orthographic similarity in recall are confounded by semantic similarities, with similar names more likely to share semantic features, because they will share STEMS (Bryan et al. 2015). While some studies have touched on semantic aspects, they have largely focussed on clinical parameters like dosage and therapeutic indication rather than the conceptual properties of the names (Lambert et al., 2002; Lambert, Chang, & Lin, 2003). Others have used too small a sample size for findings to be broadly accepted (Bryan et al., 2015). Drawing on Saussure's theory of the inextricable link between form and meaning in linguistic signs, **this thesis seeks to fill this gap in the literature by analysing both the formal and semantic aspects of generic drug names**. This is a natural development from previous research by this author on WHO naming guidelines, formal and semantic properties of INNs, and the taxonomy of INNs, and overall works towards a specialisation in the semantics of international drug nomenclature.

This chapter critically examines solutions to LASA errors within a framework (Figure 3.5). Solutions are offered at three distinct stages: pre-approval, during the medication process, and during learning and reporting. LASA error reduction strategies were presented and discussed through the lens of the hot cheese model (Thimbleby et al., 2002). All stages are important, but there is growing support for pre-approval strategies, especially those focussing on optimising drug names before they enter the market (Lambert et al. 2005; Lambert al. 2019). The third stage of learning and reporting enriches the overall body of knowledge, informing both the pre-approval stage, by delineating features that can be quantified and tweaked, and the medication process, by identifying new potential LASA name pairs that can be added to watch lists. The chapter argues that regulators and manufacturers must bear greater responsibility for preventing LASA errors at the pre-approval stage.

This thesis specifically addresses the under-researched area of LASA errors between generic drug names, specifically International Nonproprietary Names (INNs) assigned by the World Health Organization. Although LASA errors between brand names are more common, as evidenced by ISMP data, research on generic drug names has been disproportionately scarce (Basco et al., 2010; Bulfone et al., 2012; Lambert et al., 2005; Shah et al., 2017). **This thesis addresses this gap in the literature, by focussing the analysis on INNs published by the WHO and on LASA errors between generic names, rather than brand names.**

The linguistic analysis of formal and semantic features of English-language INNs is the backbone of this study, but this is accompanied by an analysis of translation and transliteration of generic names into Italian national generic names (*Denominazioni Comuni Italiane*) and Russian INNs (*Mezhdunarodnyi Nepatentovannye Naimenovanya*). **To my knowledge, this is the first study to look at this and this thesis addresses the lack of knowledge about how international drug nomenclature is presented multilingually.**

To conclude, this thesis addresses three significant shortcomings in the literature: the underexplored role of semantic features of names in LASA errors; the neglect of generic drug names in favour of commercial brand names; and the lack of research on international drug nomenclature in languages other than English. Subsequent chapters will explore the INN nomenclature by creating a taxonomic database of English-language INNs (Chapter 4), exploring the association between linguistic features and the risk of LASA error (Chapter 5) and explore the implications of these findings in other languages (Chapter 6).

Chapter 4: A taxonomic database of non-proprietary drug names in English

4.1 Introduction

Chapter 4 describes the creation of a taxonomic database of non-proprietary names in English and presents a descriptive analysis of the linguistic properties of INNs, addressing *RQ1: How can the INN nomenclature be characterised in linguistic terms?* The database serves to characterise the body of names in linguistic terms, looking at orthographic, morphosemantic, and relational properties of the names. Orthographic features pertain to the written form of names, morphosemantic features pertain to how meaning is communicated using pharmacological STEMS, and relational features pertain to how names are related to each other within the dataset. Orthographic and morphosemantic features are absolute measures, whereas relational features are relative measures, mediated by other values in the dataset.

The previous Chapter 3 reviewed the literature on LASA errors and their potential solutions, expanding upon published work in Bryan et al. (2020a, 2020b). The literature review identified a lack of focus in the literature on the drug naming stage and an undue shift in responsibility towards practitioners and away from manufacturers and naming bodies such as the World Health Organization (WHO). The core analysis seeks to identify features of INNs that act as latent conditions, increasing the probability of LASA error, in particular linguistic features (Reason, 2000). Earlier studies looking at linguistic features focus exclusively on formal aspects, especially orthographic and phonological similarity, but also neighbourhood density and frequency (Lambert, Chang, & Gupta, 2003; Lambert et al., 2001a, 2001b; Shah et al., 2017).

4.2 Primary data collection

The generic names of medications are the focus of this analysis. LASA errors between brand names outnumber those between generic names almost two to one (58% versus 33% respectively in the ISMP's 2019 list), and this seems to have led to a lack of research into generic names, with several analyses of the brand name nomenclature, but none that look solely at generic names (Basco et al., 2010; Bulfone et al., 2012; Lambert et al., 2005; Shah et al., 2017). Names in three languages were collected as primary data and compiled into a trilingual corpus.

The data analysed are cross-sectional, not longitudinal - INNs published between 1951 and 2019 are included, and ATC codes for those names were collected up to February 2021. The primary data are listed in Table 4.1.

Table 4.1 Primary data collection

Nomenclature	Type of data	N	Sourced from
English International Nonproprietary Names (INNs)	String	1046 [100%]	WHO INN lists, cross-verified on MedNet
Russian INNs (Международные Непатентованные Наименования)	String	1046 [100%]	WHO INN lists, cross-verified on MedNet
Italian DCIt (Denominazioni Comuni Italiani)	String	1046 [100%]	Sourced from AIFA, but names are designated by the Italian Pharmacopoeia
Whether or not the INN has been reported in a LASA error (this is the outcome measure)	Categorical [yes/no]	1046 [100%]	Multiple sources, including ISMP and case reports in the literature
Year in which the INN was first published in English	Continuous	1046 [100%]	WHO MedNet
Prescribing frequency	Continuous	1046 [100%]	NHS Digital, England only, ten-year average between 2009 and 2018
ATC codes assigned to each INN	String	1046 [100%]	WHO MedNet, and the ATC index

4.2.1 INNs in English and Russian

INNs in English were collected under RQ1 (see Section 1.3), and INNs in Russian were collected for analysis under RQ3. INNs are published first in English, and then translated into the other 5 official languages of the United Nations (Arabic, Chinese, French, Russian, and Spanish), and additionally Latin (SoINN, 2020b). To validate each term, the corresponding PDF list was cross-checked and the official ID of each INN was entered into the spreadsheet. This also recorded a link to an online page with details of the substance (the WHO calls this the pharmacopoeial monograph), although MedNet access is required to view this. INNs were collected from the WHO website, first collected via their biennial published lists (WHO, 2020a) and cross-checked on MedNet (WHO, 2020b). The year and decade in which the name was first designated - i.e. published in a list by the WHO - was also recorded.

Inclusion criteria for INNs were:

1. Single-word INN

2. Recommended by the WHO between 1 January 1951 and 31 October 2019 (i.e. published in lists 1 to 82 inclusive)
3. Recommended as an INN (not withdrawn or discontinued) in MedNet (WHO, 2020b)
4. Medicinal product currently authorised in the UK (listed on the Electronic Medicines Compendium and listed in the BNF).

As in the previous study, multi-word names are excluded, so that measures such as name length can be compared (Bryan et al., 2015). During the collection of primary data, multi-word names were screened for inclusion in the dataset, but they were eventually excluded, because they are not comparable to single-word names and so will skew findings. Many multi-word names use modifiers (e.g. sodium or acid) and so are no more morphologically complex than a single-word name, but will be longer and contain more syllables. Furthermore, some names include a modifier in their INN form but not in other forms, such as the USAN or BAN, and in many instances the modifier is dropped in usage.

Initially, 136 multi-word names were included in the dataset, with only one name (<1%) associated with a reported LASA error (nicotinic acid), compared with 192 single-word names (19%). The primary error source, ISMP, was explored in more detail to assess the reasons for this disparity. Of 969 LASA pairs in the ISMP list, 224 include a multi-word name (23%):

- 161 pairs involve a brand name, e.g. MS Contin - OxyCONTIN
- 31 pairs involve a common chemical name or the name for a vaccine, e.g. influenza virus vaccine; proton pump inhibitors
- 20 pairs involve the same pharmaceutical ingredient and differ only by their salt or ester, e.g. dexmethylphenidate mixed salts - dexmethylphenidate
- 10 pairs involve names that are not INNs, e.g. lithium carbonate - lanthanum carbonate
- The remaining two pairs comprise names that are both INNs but are not both in the dataset, because one is not authorised for use in the UK, e.g. ethaverine (USAN, not authorised in the UK) and etravirine (authorised in the UK).

The corpus includes names from lists 1 to 82, designated up to October 2019 and on Mednet. This is a total of 9,333 names (8,268 one-word names and 1,065 multi-word names). Only INNs that are also listed in the electronic Medicines Compendium (eMC) are included in the corpus (eMC, 2021). If listed, the medicinal product is currently authorised (and not discontinued) for marketing and human use in the UK, and is used in UK healthcare

settings. Of 9,333 names, only 1,162 returned search results on eMC; of these, 116 comprised more than one word and so were excluded, leaving 1,046 INNs in the dataset.

4.2.2 ATC codes

Most but not all INNs are assigned an ATC code by the WHO, and some have up to ten codes. Some INNs have not yet been assigned an ATC code, e.g. poloxamer and fosaprepitant. Empty ATC fields in MedNet were cross-checked and populated manually using the ATC index on or before February 2021 (WHO, 2021a).

4.2.3 Italian DCIt (Denominazioni Comuni Italiane)

Names in Italian were collected for analysis under RQ3 (How can Russian-language and Italian-language pharmaceutical nomenclatures be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?). Italian DCIts were taken from WHO MedNet and the Banca Dati Farmaci, a database managed by the Italian Pharmacy Agency (AIFA), with related information provided and monitored by the European Medicines Agency (EMA). Attempts to gain access to an online European Pharmacopoeia or Italian Pharmacopoeia were unsuccessful and are outlined in the Communications log in Appendix section A1. AIFA is challenging to use: it is not possible to search using the English name, so for names that do not have the same form in Italian, you need to guess the generic name from calquing, and thus many names will be missed, simply because they have not been guessed correctly. For example, methoxyflurane is translated as metossiflurano, which would not be found by searching for 'meth-' or 'methoxy'. This prolonged the time taken to build the Italian corpus, because broader searches had to be conducted using just the first two or three letters of a generic name and results were hand searched. It also means that some Italian names may be missing from the dataset because their correct forms couldn't be guessed. Related cognate syllables (such as English *meth-* and Italian *met-*) were noted and used to identify Italian equivalents of other names using those syllables. Google translate suggested some forms for names that were not listed on AIFA but since these forms were not on the official list they were not used.

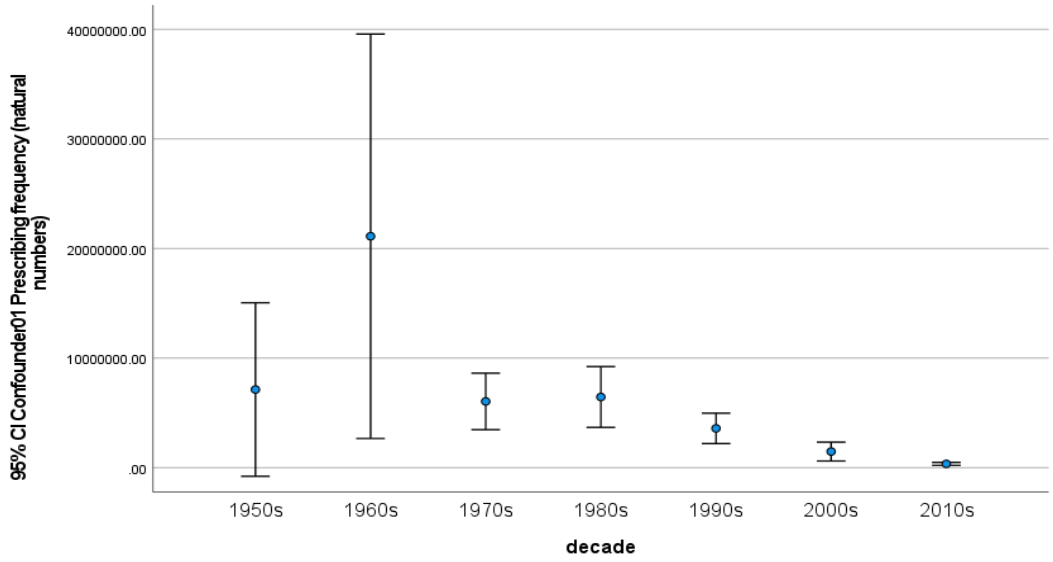
4.2.4 Prescribing frequency

Prescribing frequency data were used as a proxy for linguistic frequency (i.e. how often a word is used in the language). Prescribing frequency in the dataset is the sum of two sources of data:

1. Community prescribing - items dispensed in the community in England and Wales (NHS Digital, 2019a);
2. Hospital prescribing - hospital stock data, items dispensed in hospitals in England (NHS Digital, 2019b).

Communications with the data team at NHS Digital confirmed that these are two mutually exclusive datasets, with no possibility of overlap or data duplication. A single-year snapshot was taken, since a multi-year average might be skewed in favour of medicinal products that have been authorised for longer. 2019 was chosen: it covers the newest INNs in the dataset, but should represent a 'normal' year in comparison to the anomalous pandemic years that followed. Welsh hospital prescribing data is not published through NHS Digital for 2019, but this should have minimal implications to the variable since the population of Wales makes up around 5% of the population of England (Office for National Statistics, 2023). Names occur in multiple formulations, and sometimes in combination with other products. A python script was used to aggregate the number of prescription items for each INN. For example, *Abacavir 600mg / Lamivudine 300mg tablets* counts towards the totals for both abacavir and lamivudine. The range for this variable was broad (spanning 1.19 billion items) and exhibited extreme kurtosis (571) and skewness (21.7). To address the issue of distribution, natural numbers were transformed by adding 1 to all figures and taking the log to base 10. This yielded an interquartile range of 9.07, with kurtosis of 159 and skewness of -1.16. There is a wide range displayed, from a significant number of INNs prescribed only a few times per year (e.g. temsirolimus, fremanezumab, vindesine, nandrolone) to others with tens of millions of items prescribed every year (e.g. paracetamol, lactulose, metronidazole). Drugs named in the 1950s and 1960s have some of the highest prescribing frequencies, such as paracetamol, amitriptyline, and salbutamol, and the average prescribing frequency drops from then (see Figure 4.1).

Figure 4.1 Prescribing frequency, by decade



4.2.5 Names reported as having been involved in a LASA error

The data on names reported in LASA errors will be used as the outcome measure for *RQ2: Are any of the linguistic properties described in RQ1 correlated with the occurrence of LASA error?* A list of error pairs of names is collated that that have been reported as having been involved in a LASA error. If an error pair fulfilled certain inclusion criteria, both names in the pair would be given a positive value (“1”, or “yes”) for the *Error* variable. Thus, the *Error* variable is not pairwise but assigns a value to each name based on whether it has been involved in a LASA error or not.

Error pairs were collected from multiple sources:

1. The Institute for Safe Medication Practices (ISMP), a Canada-based non-governmental institution promoting global safety in medication practice. They collect and collate error incident reports and case studies and disseminate news articles and best practice for healthcare practitioners. They maintain a list of names (brand and generic) that have been confused with each other. These are based on events voluntarily reported to ISMP through either the ISMP national medication errors reporting programme (ISMP MERP) or the ISMP national vaccine errors reporting programme (ISMP VERP). They also publish a list of FDA-approved and ISMP-recommended Tall Man-lettered names (ISMP, 2019). The frequency of errors is not reported.

2. The Food and Drug Administration (FDA), a US-based public institution. They maintain a list of names reported to be involved in LASA errors via the national Medication Errors Reporting Programme (MERP, US Food and Drug Administration, 2001).
3. Case reports in the academic literature. As part of the literature search in Chapter 3, case reports of LASA errors were captured and collated. The search carried out as part of Chapter 3 (strategy given in Appendix section A2) also captured case reports of medication confusion errors. Searches were run on literature specific to Russia and Italy, but no reports were identified.
4. Lists of LASA error pairs in the academic literature, with unknown origin - many papers provide sample lists of LASA error pairs from reporting programs, and these were collated where possible.
5. LASA error pairs that have been automatically detected in clinical settings using electronic health data, here from Lambert et al. (2019)
6. LASA error pairs that have been intercepted in clinical settings using indication alerts, here from Basco et al. (2016) and Galanter et al. (2014).

Both names in the error pair had to fulfil these inclusion criteria:

- A generic drug name that is identical to the INN (most British Approved Names [BANs] or United States Adopted Names [USANs] are the same as the INNs)
- Included in the corpus for this study (i.e. published between 1953, the date of the first list, and 2019, and authorised for use in the UK)

Although many papers refer to lists of LASA error pairs, these could not be accessed online, and links were broken or out of date. The bulk of EPs are sourced from the ISMP, but some are from case reports or provided as part of analyses in the academic literature. Only the EPs provided by Basco et al. (2016) give the number of people subjected to an error (e.g. azathioprine was given instead of the intended azithromycin to 63 people).

Most of the pairs are reciprocal, indicating that similarity is an inherent property in the pair, not a common name overriding a less common one. Many pairs are listed in multiple sources - after duplicates were removed, 162 discrete error pairs remained with 200 unique INNs.

The list of error pairs is given in full in Appendix section A6 and comprises:

- 93 reciprocal pairs from the ISMP list (2019), each name was confused with the other name

- 26 reciprocal pars from the MER list
- 3 pairs (1 reciprocal, 2 non-reciprocal) from case reports in Chisholm & Klanduch (2000) and Cohen (2014, 2015)
- 6 reciprocal pairs from McNulty & Spurr (1979)
- 78 pairs (16 reciprocal, 62 non-reciprocal) from Basco et al. (2016)
- 8 pairs (2 reciprocal, 6 non-reciprocal) from Lambert et al. (2019)
- 17 pairs (all non-reciprocal) from Galanter et al. (2014)

The majority (94.6%) of the names that are involved in reported LASA errors are paired with just one, two, or three other names (e.g. sumatriptan, confused with saxagliptin, somatropin, and zolmitriptan). One name (lorazepam) has been confused with four other names (loratadine, oxazepam, alprazolam, and clonazepam) and another (clonazepam) has been confused with six other names (olanzapine, alprazolam, clobazam, clozapine, lorazepam, and clonidine).

There are several limitations to the data on reported LASA errors. Reported errors might be very different in nature - some were automatically detected, some were intercepted and did not reach patients, some caused serious harm and others were benign. Logically, there will be a skew in the data towards errors (either potential or actual) that have serious consequences, as a healthcare professional will reason that these are of interest to colleagues and will be more likely to report the pair. Of course, these pairs are also of more interest in the present study. Frequency is not typically stated in these sources, so there is no way to distinguish between pairs of names that have been confused hundreds or thousands of times, and pairs that have only been confused once. Errors between two generic names are in a minority in lists of LASA errors, and most pairs are reciprocal (e.g. hydralazine has been mistaken for hydroxyzine, and vice versa).

4.3 Orthographic features

In this section, the choice of orthographic variables is defined, explained, and justified, and their distribution in the data is presented using descriptive statistics. The taxonomic database contains four orthographic variables: number of characters (Orth-01), number of syllables (Orth-02); whether the INN contains any letters or groups of letters that are prohibited in the WHO naming guidelines (Orth-03); and the INN's Scrabble score, normalised by number of characters (Orth-04). Orthographic variables were measured for the full dataset (N=1046 names), because they do not rely on information from the WHO Stem Book or on any information about STEM taxa.

4.3.1 Selecting and measuring orthographic variables

The present analysis replicates and expands a previous analysis that explored formal distributional features of United States Adopted Names (USANs) and US brand names, including number of characters and syllables (Lambert et al., 2001a). It also updates and complements previous research of a larger dataset of INNs (n ~7000), including those names for drugs not currently approved for use in the UK (which are excluded in the present study), which looked at a number of orthographic and semantic features of INNs, including prohibited letters and digraphs in INNs as a way of measuring compliance with INN guidelines (Bryan et al., 2015).

The WHO prohibits the use of certain letters to reduce orthographic depth (although this is not stated explicitly), as well as to ensure that names are understood in other languages - for example, the letters <j> and <y> are not used in many languages that use the Roman alphabet. This is explored more in Bryan et al. (2015). In pure linguistic terms, it strives for orthographic shallowness, i.e. a close correspondence between letters (graphemes) and sounds (phonemes). Here, prohibited letters are measured as a proxy for orthographic shallowness, with higher scores associated with lower association between the written and spoken form of the INN. The methods used to analyse orthographic measures are described in Table A5.1.

The Scrabble score for each INN was also calculated, to be used as a proxy measure for character frequency in the name, and a measure of the orthographic complexity of a name. This was calculated as the sum of the score for each letter in the English-language version of the board game Scrabble. For example, abacavir scores 15 (1+3+1+3+1+4+1+1) and atezolizumab scores 34 (1+1+1+10+1+1+1+10+1+3+1+3). This score is calculated as if the name is entered on its own, on non-premium squares. The scoring method in the game Scrabble is based on the frequency of letters in a particular language - rare letters are assigned higher scores than common letters. Scores and boards vary by language - in English-language Scrabble, there are 42 vowel tiles, but in Italian there are 57 (in both cases A, E, I, O, and U). There are no letters K, W, X, or Y in the Italian game. Assigning a Scrabble score illustrates the complexity of each name. Some of the Scrabble scores in the INN dataset are high, probably because drug names are in part derived from scientific terms, such as oxygen, phenol, benzene, diazepine, and Greek letters such as *kappa*, which contain letters with higher scores. The Scrabble score normalised by dividing by the number of letters in the name to account for variance in name length.

4.3.2 Distribution of orthographic variables

Descriptive statistics are given in Table 4.2 and discussed in turn in the subsections below. Measures of central tendency are given for the three continuous variables (Orth-01, Orth-02, and Orth-04) and frequencies are given for the categorical variable (Orth-03).

Table 4.2 Summary of the distribution of orthographic variables in the taxonomic database

ID	Variable	Type of feature	Type of data	N valid cases [% of total cases]	Descriptive statistics / frequencies			
					Minimum	*Maximum*	*Mean*	*Std deviation*
Orth-01	How many characters does the INN contain?	Orthographic	Continuous	1,046 [100%]	5	23	10.60	1.87
Orth-02	How many syllables does the INN contain?	Orthographic	Continuous	1,046 [100.0%]	2	9	4.17	0.80
Orth-04	What would the INN score if played in a Scrabble game?	Orthographic	Continuous	1,046 [100.0%]	1	4	1.72	0.43
					Yes [% valid cases]		*No [% valid cases]*	
Orth-03	Does the INN contain prohibited letters or groups of letters?	Orthographic	Categorical [yes/no]	1,046 [100.0%]	192 [18.4]		854 [81.6]	

4.3.2.1 Orth-01, number of characters; and Orth-02, number of syllables

The average INN has 10.6 characters and 4.1 syllables (examples are mesalazine and salmeterol). This is consistent with United States Adopted Names (USANs) in Lambert et al. (2001a), with figures compared in Table 4.3. The modal numbers are the same, but INNs are longer on average and contain more syllables. These analyses are two decades apart, with this dataset including names approved up to 2019, but the numbers have remained relatively stable. Density plots are given in Figures 4.2 and 4.3.

Table 4.3 Descriptive statistics for INNs (n=1,046), Orth-01 and Orth-02

	Characters (Orth-01)		***Syllables (Orth-02)***	
	This analysis (n=1,046)	**Lambert et al., 2001 (n=8,712)**	**This analysis (n=1,046)**	**Lambert et al., 2001 (n=8,712)**
Mean	10.60	10.46	4.17	4.02
Median	10.00	10.00	4.00	4.00
Mode	10.00	10.00	4.00	4.00
Min	5.00	3.00	2.00	1.00
Max	23.00	26.00	9.00	-
1st quartile	9.00	-	4.00	-
3rd quartile	11.00	-	5.00	11.00
Range	18.00	-	7.00	-
SD	1.87	1.97	0.80	0.83

Figure 4.2 Density plot for Orth-01, number of characters

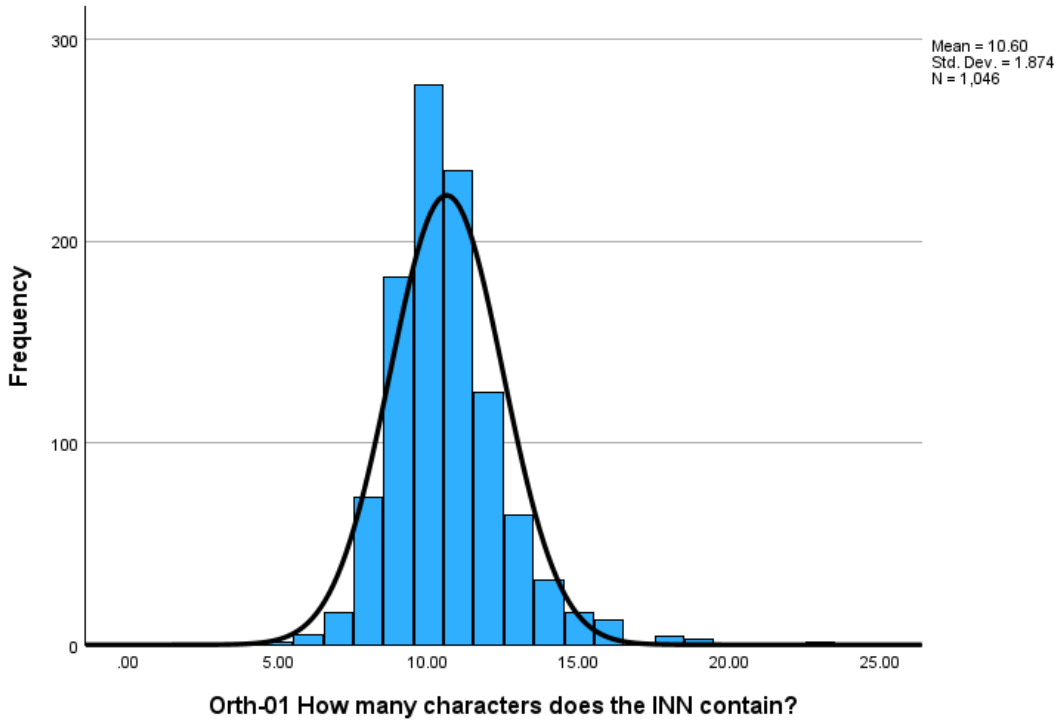
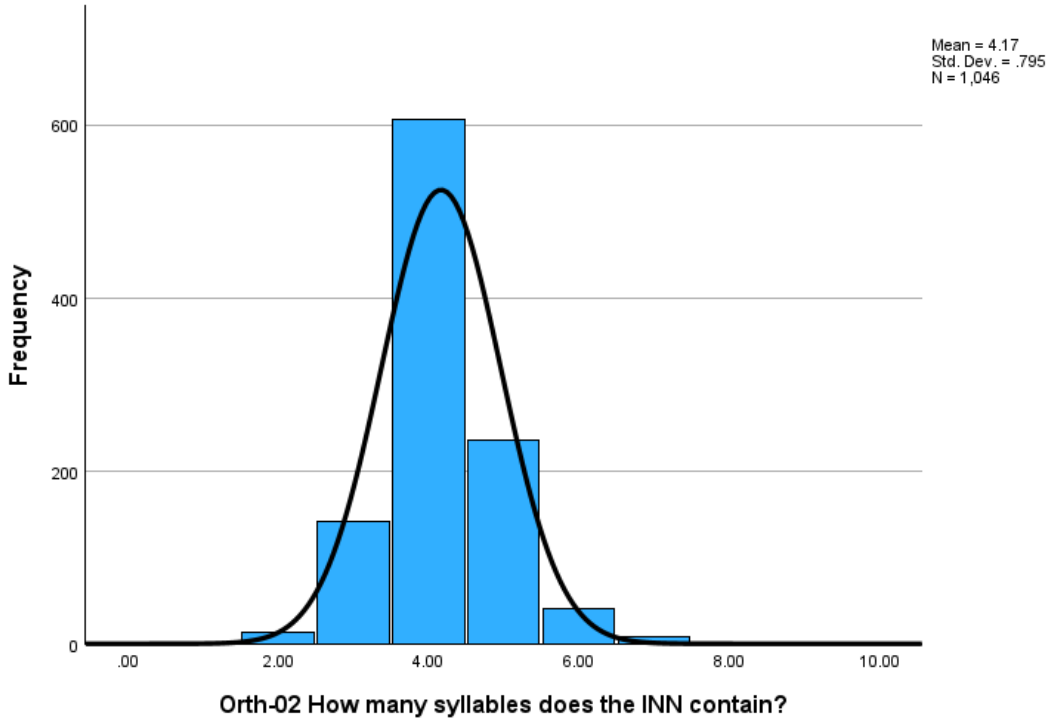


Figure 4.3 Density plot for Ortho-02, number of syllables



These findings are consistent with an earlier study led by this author (Bryan et al., 2015), in which a larger corpus of one-word INNs were measured. In that study, mean length was 10.54 (here it is 10.60), with standard deviation 1.73 (here 1.87). It appears that although

this corpus only includes names for drugs that are authorised for use in the UK, in relation to name length, it is representative of the corpus as a whole. The previous analysis did not measure the number of syllables. The number of characters and syllables was high in the early stages of the INN programme (fifties and sixties) but dropped sharply from the seventies and has remained steady since then.

4.3.2.2 Orth-03, presence of prohibited letters

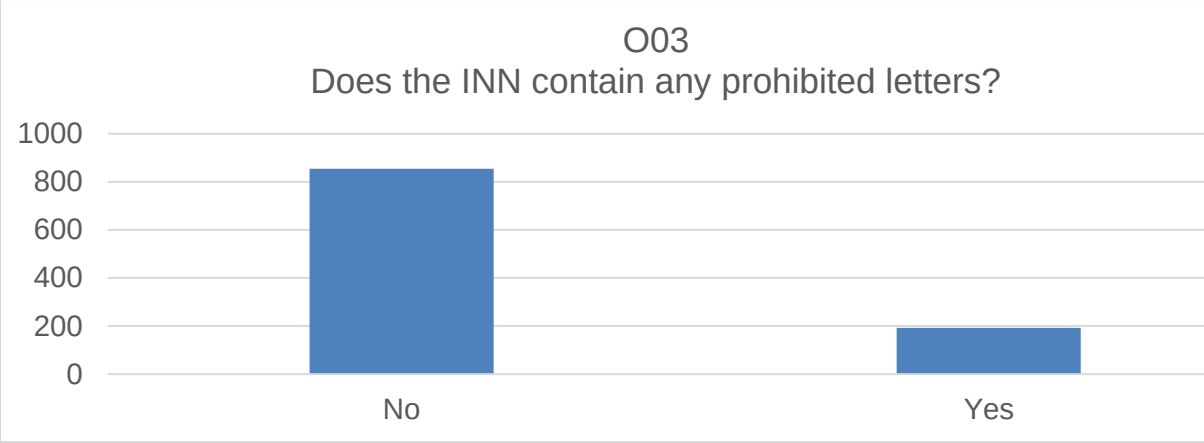
The number of prohibited letters occurring in INNs was used as a proxy measure for orthographic depth, i.e. the correspondence in the corpus between the number of sounds and the number of letters, and is summarised in Table 4.4. Almost a fifth of all names (18.36%, 192/1046) contain at least one prohibited letter, and only one INN contains five prohibited letters or groups of letters (phenoxymethylpenicillin, with <ph>, <y>, <th>, <y>, and duplicate <ll>). Four INNs contain four prohibited letters or groups of letters (dihydrotachysterol, hydrochlorothiazide, hydroxychloroquine, and trihexyphenidyl). Similar to previous findings from an analysis of the full INN nomenclature, the incidence of using prohibited letters has decreased steeply since the 1960s, despite the number of INNs published each year remaining relatively steady (see graph in Appendix section A9).

The most common duplicate letters were ll (in 14 INNs). None of the duplicate letters fall across the boundary of two formatives, such as in English *bookkeeper*. As identified previously (Bryan et al., 2015), many of the STEMS that names are required to include use a prohibited letter so their use is unavoidable, e.g. -cillin. It's also worth noting that in English, a duplicate letter does not change the pronunciation of the sound represented by a single letter, although it may change the stress pattern of the word or the pronunciation of adjacent vowels. In other languages, however, a duplicate letter may represent a different sound, e.g. Spanish *llamar*, 'call' (pron. /ʎaˈmaɾ/) or Welsh *llew*, 'lion' (pron. /ɬɛʊ̯/).

Table 4.4 Prohibited letters and groups of letters in INNs, Orth-03

Prohibited letter(s)	***Frequency***	***INNs affected***	***Examples***
y	114	100	acetylcysteine [1965], moxisylyte [1966]
th	45	45	sulfamethoxazole [1966], thalidomide [1962]
h (not including ph or th)	39	35	betahistine [1965], hyetellose [1999]
ph	26	24	epinephrine [1967], fluphenazine [1962]
k	21	21	dexketoprofen [1994], montelukast [1996]
duplicate letters	20	20	benzylpenicillin [1985], desmopressin [1975]
oe	2	2	moexipril [1989]
ae	0	0	-
TOTAL	267	192	
			*Number of unique INNs with at least one prohibited letter

Figure 4.4 Bar chart for Orth-04, presence of prohibited letters



4.3.2.3 Orth-04, normalised Scrabble score

Scrabble scores, using the English-language board, were taken for each INN, and normalised by dividing the total score by the number of characters. Descriptive statistics are given in Table 4.5. The mean Scrabble score for the INNs was 18.38 (in general language, words such as ZEBRAS and JACKPOT get this score), with one name attracting a score of 50 (phenoxymethylpenicillin) and another a score of 6 (serine). The normalised Scrabble score has remained steady over the last seven decades of INNs. Density plots for the raw Scrabble score and the normalised Scrabble score in Figures 4.5 and 4.6.

Table 4.5 Descriptive statistics for one-word INNs (n=1,046)

	Scrabble score for INN	Normalised scrabble score
Mean	18.38	1.72
Median	17	1.63
Min	6	1.00
Max	50	3.50
Range	44	2.50
SD	6.08	0.43

Figure 4.5 Density plot for Scrabble score

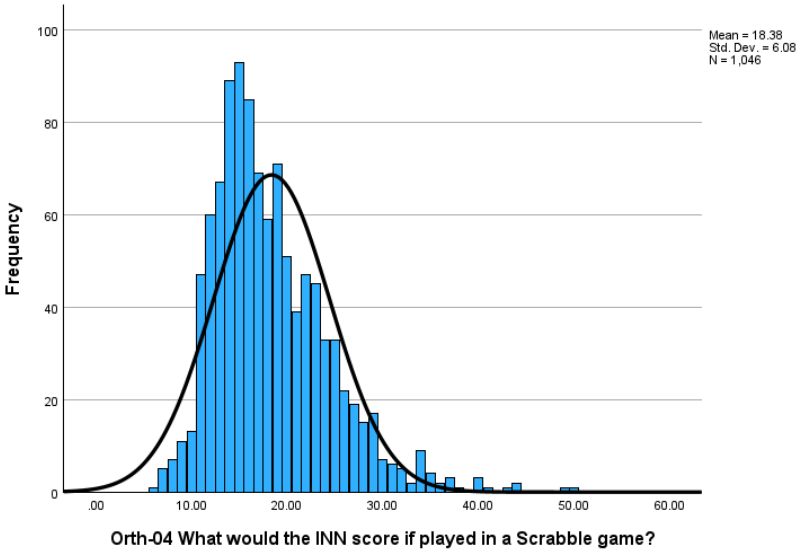
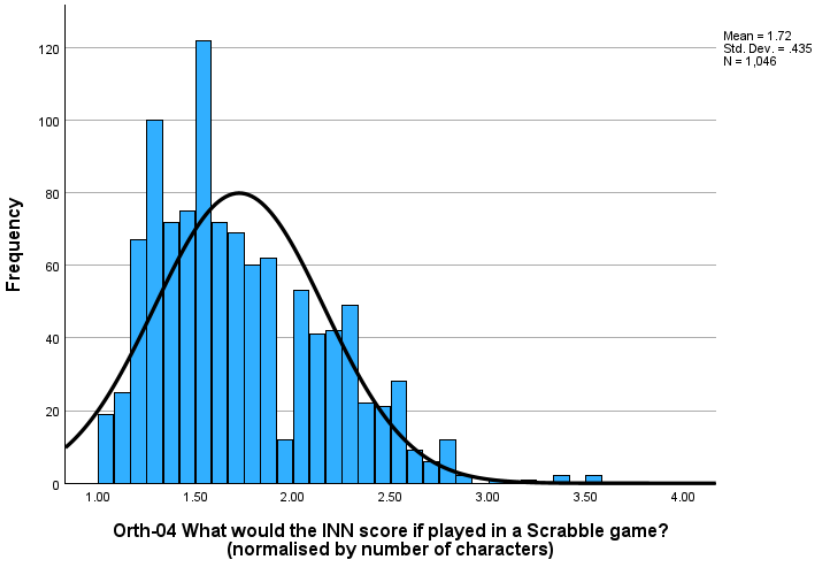


Figure *4.6 Density plot for Orth-4, normalised Scrabble score*



4.4 Morphosemantic features

Since this thesis looks at linguistic features of INNs through the lens of Saussure's linguistic sign, the focus of the analysis is on the interaction between form and meaning, and how meaning is communicated by combining STEMS (word parts). Thus, morphosemantic features form a core component of this analysis. A range of morphosemantic features are explored to identify ways in which names communicate meaning and the extent to which they may be described as semantically transparent.

4.4.1 Selecting and measuring morphosemantic variables

Eight morphosemantic measures were explored. Design of the first five variables is driven by observations in the WHO Stem Book, e.g. compliance, presence of STEMS and SUBSTEMS, variant forms, and incidental STEMS (WHO, 2021b). The Stem Book was consulted as the 'official guidance' of STEM usage in INNs, and several morphosemantic measures in this database are based on information in the WHO Stem Book. For example, Morph-01 measures compliance with naming guidelines by whether the Stem Book states the INN is compliant. Morph-04 measures the extent of variant STEMS by whether the Stem Book marks a particular STEM as a secondary or tertiary form of the 'standard' pharmacological STEM and Morph-05 measures the number of STEMS that are used incidentally, without belonging to a particular pharmacological group. FREEFIXES (Morph-06) and taxon type (Morph-08) were selected for analysis to continue previous research (Bryan et al., 2015), and looking at headedness (Morph-07) was prompted by its being a key concept in the literature on semantic transparency (e.g. Schäfer, 2018).

The Stem Book provides the necessary information for the analysis of INNs' use of pharmacological STEMS. Where information on STEMS is missing or names are deemed non-compliant, full analysis of these names is not possible and thus some variables will have missing values. Certain morphosemantic variables are used to define analysis subgroups. For example, names that are deemed compliant (Morph-01) are analysed for Morph-02 through Morph-08, because the STEM information provided in the WHO Stem Book is reliable. In turn, names that have at least five members belonging to the same pharmacological STEM family ('taxon') are analysed for Morph-07 and Morph-08, because they propose judgements about behaviour of the taxon, e.g. about its headedness or complexity. The inclusion criteria and the number of missing values for each variable is given in Table 4.6 below.

4.4.1.1 Morph-01, compliance with WHO guidelines

To measure compliance (Morph-01), the Stem Book was used (WHO, 2021b). This gives guidance on how STEMS are to be used in INNs and gives an overview of most names, but not all. The Stem Book states:

> *"For each stem, INN have been classified as:*
>
> *(a) names in which the preferred stem has been used in accordance with its definition;*
>
> *(b) names in which the preferred stem has been used but not in accordance with its definition;*
>
> *(c) names which belong to the same group of pharmaceutical substances and in which no preferred stem has been used"*
>
> Stem Book (WHO, 2021b:2)

In other words: names classified as (a) use the correct STEM; names classified as (b) use the incorrect STEM; and names classified as (c) are missing the correct STEM. When parsing each name using the Stem Book (WHO, 2021b) it was noted which of (a), (b), or (c) the name belongs to, and names in (b) or (c) were excluded from morphosemantic analysis because use of STEMS was a core element of this analysis and names using the wrong STEM or not using a STEM at all would risk skewing findings. Furthermore, during data collection it was noted that there were very few names classified as (b) or (c) so the decision to exclude them did not significantly reduce the size of the database.

4.4.1.2 Morph-02, presence of STEMS; Morph-03, presence of SUBSTEMS

INNs show their meaning and semantic interrelationships through a taxonomic system of STEMS. These are word parts that show pharmacological grouping, such as vir for antivirals. The INN Stem Book outlines the classification system, and lists the names occurring under each STEM, further noting when names diverge from official guidance and contain the incorrect STEM or are missing a STEM. Each INN was morphologically parsed (broken down into the smallest possible units with meaning) into STEMS and so-called 'random [or fantasy] PREFIXES'. To count the presence of a STEM and/or a SUBSTEM, a list of STEMS was extracted from the Stem Book (WHO, 2017a). The official classification contains 324 unique STEMS, and a further 296 STEMS were extracted by hand searching the Stem Book and noting variant spellings (allomorphs), and additional STEMS and SUBSTEMS. Stems that differ only by their position in the word were counted as unique, as the previous study by this author showed that STEM position can influence meaning. For example, if mer is a PREFIX (e.g.

merbromin) or an INFIX (e.g. thiomersal), it denotes 'mercury-containing drugs, antimicrobial or diuretic'; if mer is a SUFFIX (e.g. carbomer), it denotes 'polymers'.

4.4.1.3 Morph-04, presence of variant STEMS; and Morph-05, number of incidental STEMS

Some STEMS have a ‘variant’ form, provided in the Stem Book, such as official STEM -azosin with variant STEM -uzosin in alfuzosin. These were manually noted when parsing each INN according to the STEM and SUBSTEM used.

To count the incidence of incidental STEMS, the ‘correct’ STEM taxon was confirmed manually and other STEMS present were noted as ‘incidental’. An incidental STEM is an official STEM, as listed in the Stem Book, but one that is only in the INN incidentally, and does not carry meaning, e.g. norepinephrine incidentally contains the STEMS -ine, -pin, and -pine without being part of these pharmacological groups. Where a name contains two STEMS from separate STEM taxa, the STEM that appears more frequently in the dataset was assigned to the name, e.g. alprostadil contains prost and -dil, but was recorded under prost.

4.4.1.4 Morph-06, presence of FREEFIX STEMS; and Morph-07, right-headedness

Previous research defined FREEFIXES, and they are measured here in this dataset (Bryan et al., 2015). The Stem Book uses hyphens to indicate the position of the STEM inside the INN - for example, in cef- the hyphen indicates that the STEM appears at the beginning of the name, -ster- indicates that it appears in the middle of the name, and "-golide" at the end of the name. When the STEM is listed with no hyphen (e.g. vir or pred), the names within the STEM taxon contain that STEM in any position, thus a lack of hyphens is assumed to mean the STEM is a *FREEFIX* (as first defined in Bryan et al., 2015). The Stem Book states: "where a STEM is shown without any hyphens it may be used anywhere in the name" (WHO, 2021b). When manually parsing INNs using the Stem Book, the official position of STEMS was noted (as either start, mid, end, or FREEFIX), and compared with the actual positions of STEMS.

Headedness is a useful concept for analysing neoclassical compounds that can be parsed into morphological components, because it is a marker of semantic transparency (right-headed compounds are more ‘typical’ in English thus more transparent). STEMS and SUBSTEMS were classed by their position in the name, and divergence from this was looked at more closely. To measure headedness, INNs were classified as being either:

- right-headed (showing their pharmacological grouping by placing the STEM at the end of the name, with optional further categorisation with a SUBSTEM to the left);
- left-headed (showing their pharmacological grouping by placing the STEM at the start of the name, with optional further categorisation with a SUBSTEM to the right);
- neither (STEM showing pharmacological grouping appears in different positions in the name, or in the middle of the name, and SUBSTEMS either change the form of the STEM or are not attached to the STEM).

In this sense, a STEM that is neither left-headed nor right-headed might have a STEM that appears in the middle of the name, with a SUBSTEM at the end. It may be exocentric, with no STEM but only the SUBSTEM.

4.4.1.5 Morph-08, taxon type of the INN STEM

To classify STEM taxa into the types outlined earlier, the structure of the STEM taxon was also noted - whether there was a single STEM, or a combination of STEM + SUBSTEMS, or STEM + multiple SUBSTEM systems. In a book chapter based on the preceding MA thesis (Bryan, 2015), the INN STEM taxonomy was classified into four types (single-level, multi-level, complex, and irregular). This typology was initially based on a small dataset of 71 INNs and a narrative description of the dataset and created inductively by exploring features of the dataset. In this analysis, taxon type is used as a variable to distinguish various shapes of taxon, and the typology is further specialised by exploring a broader suite of measures, such as the presence of FREEFIXES, the presence of STEMS and SUBSTEMS, and headedness.

4.4.2 Distribution of morphosemantic variables

These measures were conducted on a subset of the dataset, and the sample sizes are outlined in Table 4.6. All variables are categorical with values “yes” and “no”, with the exception of Morph-05, the number of incidental STEMS, which is continuous. Measures of central tendency are given for the continuous variable (Morph-05) and frequencies are given for the seven categorical variables.

Table 4.6 Summary of the distribution of morphosemantic variables in the taxonomic database

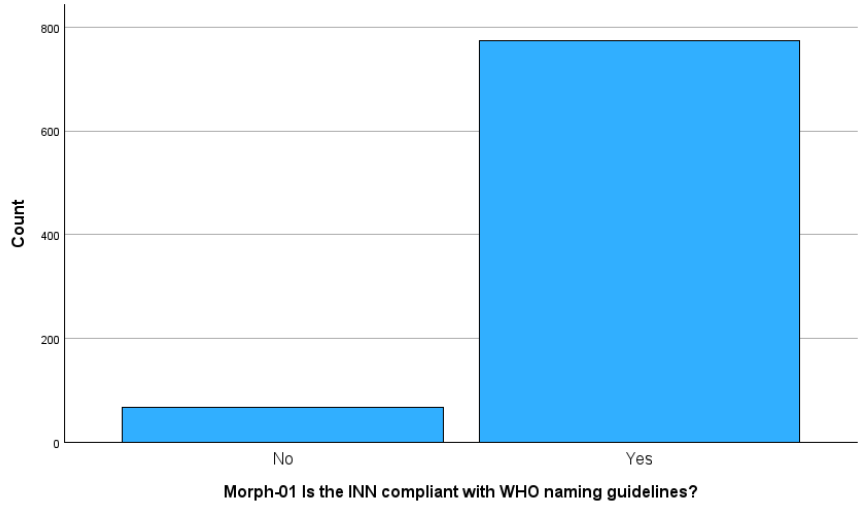
ID	Variable	Type of feature	Type of data	N valid cases [% of total cases]	Inclusion criteria	Descriptive statistics / frequencies			
						Minimum	Maximum	Mean	Std deviation
Morph-05	How many incidental STEMS appear in the INN?	Morphosemantic	Continuous	774 [74.0%]	Names that are listed as 'compliant' in the Stem Book	0	6	0.80	1.01
						Yes [% valid cases]		No [% valid cases]	
Morph-01	Is the INN compliant with WHO naming guidelines?	Morphosemantic	Categorical [yes/no]	842 [80.5%]	Names that are listed in the Stem Book	774 [91.9]		68 [8.1]	
Morph-02	Does the INN contain a STEM?	Morphosemantic	Categorical [yes/no]	774 [74.0%]	Names that are listed as 'compliant' in the Stem Book	746 [96.4]		28 [3.6]	
Morph-03	Does the INN contain a SUBSTEM?	Morphosemantic	Categorical [yes/no]	774 [74.0%]		291 [37.6]		483 [62.4]	
Morph-04	Does the INN contain a variant STEM?	Morphosemantic	Categorical [yes/no]	774 [74.0%]		16 [2.1]		758 [97.9]	
Morph-06	Does the INN include a FREEFIX STEM?	Morphosemantic	Categorical [yes/no]	774 [74.0%]		147 [35.1]		272 [64.9]	
Morph-07	Is the INN right-headed?	Morphosemantic	Categorical [yes/no]	452 [43.2%]	Names that are listed as 'compliant' in the Stem Book and belong to a pharmacological STEM taxon with at least five names	312 [69.0]		140 [31.0]	

						Type 1 Single-level [%]	Type 2 Multi-level [%]	Type 3 Complex [%]	Type 4 Irregular [%]
Morph-08	Which taxon type does the STEM in this INN belong to? (types 1, 2, 3, or 4)	Morphosemantic	Categorical [yes/no]	452 [43.2%]	Names that are listed as ‘compliant’ in the Stem Book and belong to a pharmacological STEM taxon with at least five names	151 [33.4]	89 [19.7]	63 [13.9]	149 [33.0]

4.4.2.1 Morph-01, compliance with WHO INN naming guidelines

The Stem Book was consulted as the 'official guidance' of STEM usage in INNs. Of 1,046, 204 were not listed in the Stem Book so could not be deemed compliant to official guidance or not. The remaining 842 were listed in the Stem Book, of which 774 names (91.92%) were deemed compliant ("a: names in which the preferred STEM has been used in accordance with its definition", i.e. names that use STEMS correctly). This is illustrated in the bar chart in Figure 4.7. In a smaller sample of names in previous work, 28 of 71 names (39.44%) were either not in the Stem Book, or not listed as an "a" in the Stem Book, i.e. using the STEM/SUBSTEM correctly. This figure is a little lower here: 26.00% (272 of 1046) names are not in the Stem Book nor are listed as an "a" in the Stem Book. The probable reason for the lower figure in this larger dataset is that these names are authorised in the UK so more likely to be listed, because they are in clinical use and serve as illustrative examples. The sample is also much larger so should be more accurate.

Figure 4.7 Bar chart for Morph-01, compliance with WHO guidelines



4.4.2.2 Morph-02, presence of a STEM; and Morph-03, presence of a SUBSTEM

Of the 774 names found to adhere to official WHO naming guidelines, almost half contain a single STEM with no SUBSTEM, e.g. mycin in erythromycin, and estr in estradiol (483, 62.40%). Over a third contain a SUBSTEM, which may or may not co-occur with its 'parent' STEM (291, 27.8%). This is illustrated in the bar charts in Figures 4.8 and 4.9.

Figure 4.8, Bar chart for Morph-02

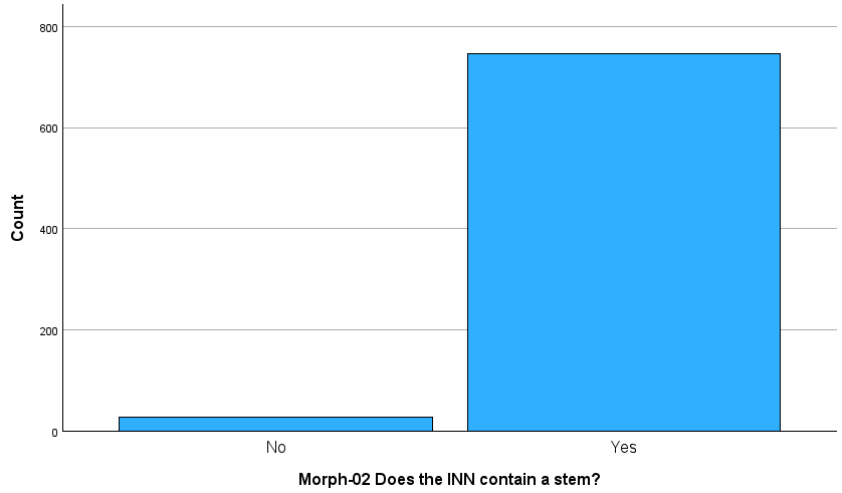
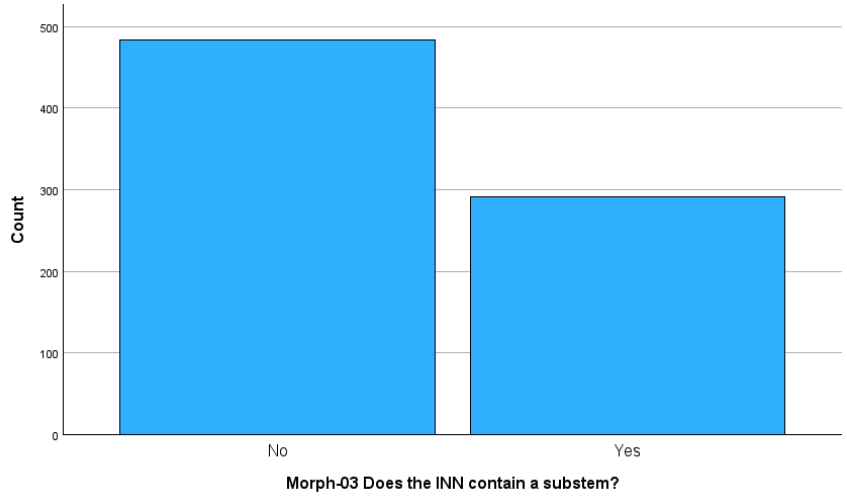


Figure 4.9 Bar chart for Morph-03



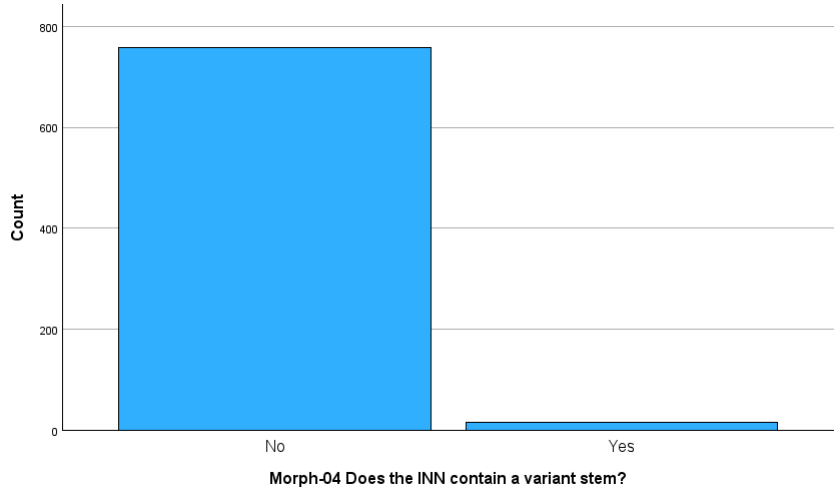
A small number (28) of these INNs with a SUBSTEM have no official 'parent STEM', but rather are a collection of SUBSTEMS under an undefined head, e.g. dopa in levodopa, and cef in cefuroxime, presented as SUBSTEMS in the Stem Book but no STEMS are given (WHO, 2021b). The phenomenon of a collection of hyponyms (SUBSTEMS) under an undefined hyperonym was first identified in Bryan et al. (2015) and seems to be a persistent problem as it also appears in this larger sample of names. This may be nothing more than a formatting error or a peculiarity of the Stem Book, and certainly displays a lack of understanding of taxonomy, but it is difficult to relate members of a STEM taxon when the main taxon is left undefined. Only 0.72% (6/830) of INNs in the sample place a STEM not according to Stem Book guidance, such as carvedilol, with end-of-name STEM -dil. Over time, STEM taxa are becoming increasingly complex with an increased use of SUBSTEMS, as

in Appendix section A11. Only a fifth of names that were approved in the 1950s and 1960s belong to a STEM taxon with a SUBSTEM, rising to more than half in the 2000s.

4.4.2.3 Morph-04, presence of a variant STEM; and number of incidental STEMS (Morph-05)

In previous research looking at the entire nomenclature of INNs, many variant STEMS were identified (Bryan et al., 2015), yet in this corpus of names for drugs authorised for use in the UK, only a handful of INNs (16/774 names, 2%) used a variant form of the STEM, such as alfuzosin (official STEM -azosin, using variant STEM -uzosin) and norepinephrine (official STEM -drine, using variant STEM -phrine). This is illustrated in the bar chart in Figure 4.10. A full list of INNs using a variant STEM is in Appendix section A10. INNs containing variant STEMS were all approved before 2000, so it is a historic and minor problem.

Figure 4.10 Bar chart for Morph-04



Variant STEMS often drop or add on an end vowel, e.g. nico- to nic-, and nitro- to nit- in ranitidine - this will ease pronunciation when the STEM is followed by a vowel so vowels are not clustered together and difficult to pronounce. The WHO guidance decrees that the name be “euphonious” (WHO 2017; WHO 2022). In other cases, it may be that an alternate form of the STEM was made official in an attempt to dispense with prohibited letters, such as with OFFICIAL STEM -fentanil and allomorph STEM -fentanyl (e.g. in fentanyl and sufentanil) and official STEM -izine and allormorph STEM -yzine (e.g. in hydroxyzine).

Many INNs contain an official STEM but only incidentally, i.e. the STEM is usually short and appears in the name, without conferring relationship to that pharmacological group. Half of INNs (50.2%, 525/1046 names) contain no incidental STEM, and almost a third (29.0%,

303/1046) contain one incidental STEM. Although the average has decreased over time (see Appendix A11), it persists in recently approved INNs. This is illustrated by the density plot in Figure 4.11.

Figure 4.11 Density plot for Morph-05.

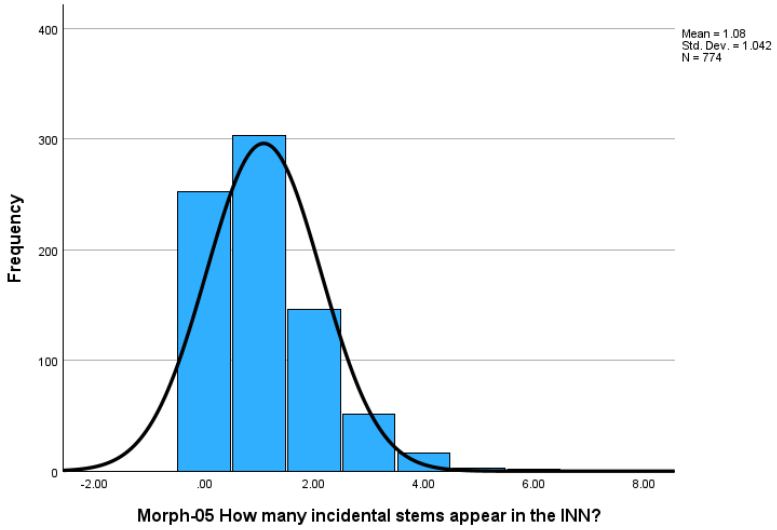


Table 4.7 gives some examples of INNs containing incidental STEMS, such as -al and -ni-. These short strings of two or three letters will inevitably appear in other INNs without acting as an official STEM. In total, 97 STEMS appear 'incidentally' in other INNs, but many only once. In itself, this does not appear to be a glaring problem, but it may be confusing for a user if they mistakenly interpret a name as belonging to the wrong taxon. This is the only continuous morphosemantic variable.

Table 4.7 Common incidental STEMS in INNs, with examples

Incidental STEM	**Frequency**	**Examples**
ine	145	vincristine, rilpivirine, doravirine, nicardipine, apraclonidine
ol	130	ketoconazole, fenticonazole, fluconazole, mestranol, gadobutrol
al	97	colecalciferol, paricalcitol, palbociclib, lenalidomide, hydralazine
ni	92	lorlatinib, fluocinolone acetonide
ac	81	brivaracetam, propylthiouracil, gimeracil, acetylcysteine, indacaterol
ox	59	duloxetine, enoxaparin sodium, sulfamethoxazole, fludroxycortide
one	41	milrinone, cyproterone, abiraterone, prasterone, paliperidone

pin	24	amlodipine, pindolol, clozapine, mirtazapine, quetiapine
ast	23	dutasteride, rivastigmine, trastuzumab, pravastatin, simvastatin
cain	13	cinchocaine, pramocaine, procaine, benzocaine, lidocaine
io	13	tioconazole, estradiol, riociguat

4.4.2.4 Morph-06, presence of a FREEFIX STEM; and Morph-07, right-headedness

The analysis of FREEFIX STEMS and headedness was restricted to names in frequent taxa only (those with at least five INNs using the STEM) – 452 names of the total sample of 1046 names. Of the 774 INNs containing an official STEM and compliant with the WHO Stem Book, 147 contain a FREEFIX STEM. This is illustrated in Figure 4.12 and Table 4.8 gives some examples. A FREEFIX STEM can appear anywhere in the name. As the table shows, its position varies within the taxon prost is at the end in dinoprost, but in the middle in dinoprostone; calci in the middle in colecalciferol, and at the start in calcitriol. Over time the number of FREEFIXES has remained steady (Appendix A11). These STEMS are defined as FREEFIXES in the WHO guidance (the Stem Book), i.e. they can appear anywhere in the name without affecting meaning or the taxon they belong to. In practice, most tend to appear in a particular position, such as the majority of names using vir display the STEM at the end of the name, as in aciclovir.

Figure 4.12 Bar chart for Morph-06

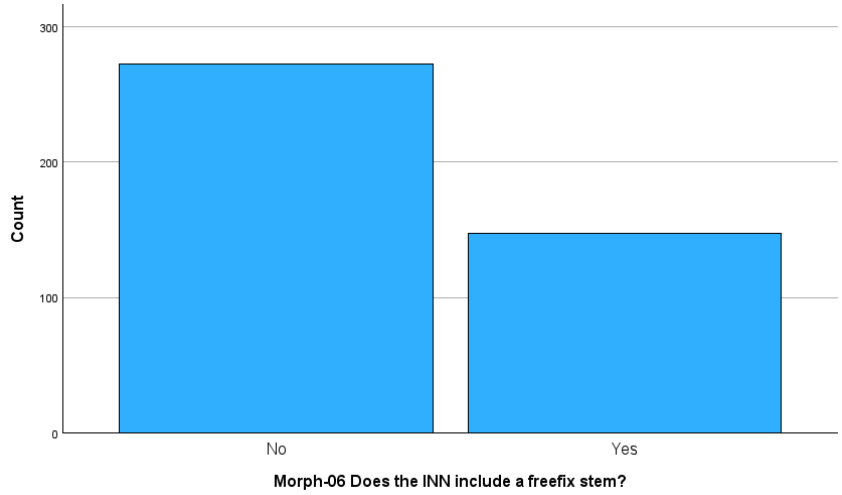


Table 4.8 Examples of FREEFIXES in INNs, with examples

FREEFIX STEM	***Frequency in dataset***	***Examples***	***Beginning***	***Middle***	***End***
vir	41	ribavirin, aciclovir, ganciclovir	0	8	33
gli	14	gliclazide, glimepiride, pioglitazone	3	11	0
stat	13	pentostatin, pravastatin, simvastatin	0	6	7
prost	12	dinoprost, carboprost, epoprostenol	0	4	8
gest	12	progesterone, medroxyprogesterone	1	10	1

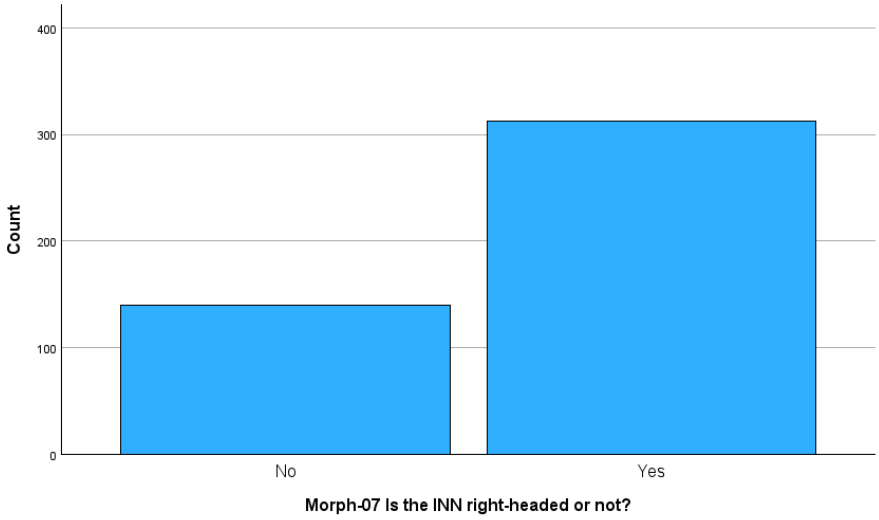
FREEFIX STEMS can appear anywhere in the name; however, most appear at the end of the name - these are deemed right-headed because their 'head', i.e. the STEM taxon that they belong to, is positioned at the end of the name, on the right (in English and other left-to-right languages). The headedness of 452 INNs is outlined in Table 4.9. The majority of these INNs are right-headed, either with one STEM only at the end of the INN, such as acriv(astine) (STEM -astine at the end), or with a STEM and a SUBSTEM, such as mor[octo](cog) alfa (STEM -cog at the end and SUBSTEM -octo- preceding it in mid position). Very few INNs (10 of 534) are left-headed, such as (nal)mefene. Those with two SUBSTEMS are partially right-headed, with the two SUBSTEMS at equal levels in the taxonomy (not a subcategory of the other), but the hyponymic STEM is placed on the right, such as in tras[tu][zu](mab). Over time (see Appendix A11) the number of right-headed INNs has increased, as has the number of INNs that are neither right- or left-headed.

Table 4.9 Headedness in INNS

Headedness	**Frequency**	**Examples**
Right-headed	312 (69.0%)	atenolol (STEM olol at the end, no SUBSTEM)
Left-headed	10 (2.2%)	cefepime (STEM cef at the start, no SUBSTEM)
Neither	130 (28.76%)	sitagliptin (FREEFIX STEM gli in mid position, SUBSTEM gliptin at end position)

Since right-headedness is typical for English compounds and for neoclassical compounds, of interest here is whether or not INNs are right-headed and this is illustrated in Figure 4.13. The majority of the INNs studied for headedness are right-headed (29.8%, 312, 69.0%).

Figure 4.13 Bar chart for Morph-07



Headedness is one aspect of the dataset that is much more heterogeneous - a large number of INNs (29%, 130/452) here are neither right- nor left-headed, and display their semantic category in a range of ways. From a closer look at names that are neither left- nor right-headed, there are two processes of building meaning with STEMS in names - (i) agglutination and (ii) replacement. In agglutination, the head (i.e. the STEM) appears in the middle of the name - usually because it is a *FREEFIX* (a term first coined in Bryan et al., 2015) - and SUBSTEMS are attached to the STEM in various positions. As neoclassical compounds, they are still endocentric as they display a head within the name, but it is not easily parsed as it appears in the middle of the name, or they appear in various positions that are unpredictable to users. In replacement, the SUBSTEM replaces the STEM and so it is impossible for the user to parse the name. This sometimes includes a change in position, or a negative SUBSTEM (SUBSTEM -ur from STEM -uridine). These compounds no longer contain a head, but a hyponym of the head, so they are transformed into exocentric neoclassical compounds. Examples are given in Table 4.10.

Table 4.10 SUBSTEMS, agglutination and replacement

Features of headedness	***SUBSTEM***	***Examples, parsed into STEMS and SUBSTEMS***	***Number of INNs (a)***	***Number of INNs (b)***
Agglutination of a SUBSTEM to a FREEFIX STEM				
	No SUBSTEM	repa(gli)nide, deso(gest)rel	32	34
	SUBSTEM attaches to beginning of STEM:	levonor[g](estr)el, me[g](estr)ol	20	20
	SUBSTEM attaches to end of STEM:	lina[(gli)[ptin], sulfa[(sal)[azine]		
	SUBSTEM attaches as a wrapper around STEM:	flu[va](stat)[in], rosu[va](stat)[in]	3	1
Replacement - SUBSTEM replaces the STEM, sometimes changing position				
	SUBSTEM in the same position:	lami[vudine] (replacing STEM -uridine), sot[alol] (replacing STEM -olol)	13	13
	SUBSTEM changes position:	[ni]car[dipine], [ni]fe[dipine] (both replacing STEM nitro-)	3	5
	Negative SUBSTEM:	tegaf[ur] (replacing STEM -uridine)	1	1

4.4.2.5 Morph-08, taxon type of the INN STEM

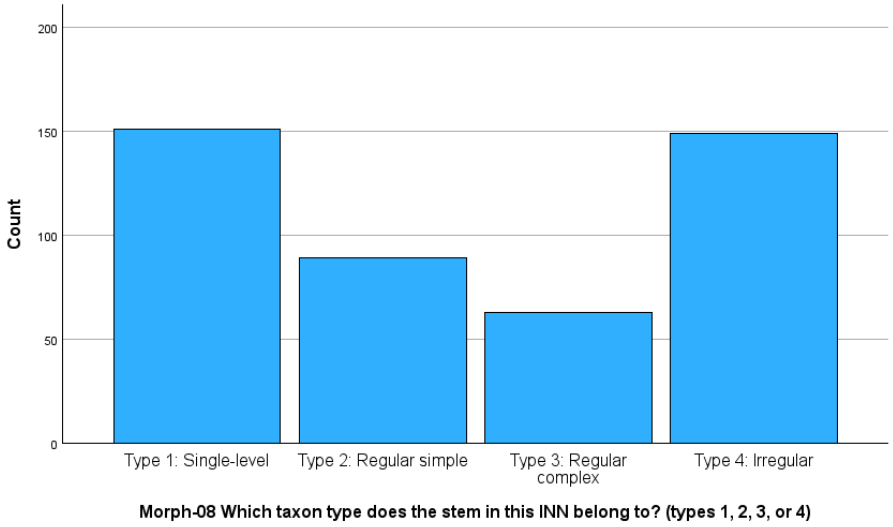
Typology of taxa, as presented earlier in the chapter relates closely to headedness, in that taxa with a simple right-to-left display of taxonomy are deemed ‘regular' since they are predictable for the user. This typology was first created in Bryan (2015) when looking at a much smaller sample of names in 2015. Here, the corpus of names is bigger and headedness is analysed for 452 names. The typology will be assessed to take into account the results from this larger analysis. In brief:

1. Single-level taxa - right-headed, a STEM but no SUBSTEM
2. Multi-level taxa - right-headed, STEM and SUBSTEM
3. Complex taxa - right-headed, STEM and multiple SUBSTEM systems
4. Irregular - not right-headed STEM, containing a variant or FREEFIX STEM, or a collection of SUBSTEMS without a hyperonymic STEM

Most of the right-headed names (78.2%, 273/349) belong to type 1 or type 2 – single-level taxa or multi-level taxa, respectively. These are taxa with only one STEM, or with a simple

taxonomic hierarchy of STEMS with hyponymic SUBSTEMS, such as monte([luk]ast) for antiasthmatics. The remaining right-headed names (76/349, or 21.8%) are predominantly monoclonal antibodies, under type 3 for complex taxa, such as tras([tu][zu]mab) emtansine.

Figure 4.14 Bar chart for Morph-08



The emergence of left-headedness in some names in the corpus is interesting, as this is a smaller proportion of names so was not identified in previous analyses. These names would belong under type 4 (irregular) as they do not show their semantic category in a predictable way – the majority of names are left-headed, and there is no mechanism to prompt the user to switch to a left-to-right interpretation of the name. Thus type 4 should be widened to include names with a left head, such as FREEFIX STEM vin- in (vin)blastine and (vin)cristine. Type 4 is also widened to include all names with a FREEFIX STEM, as they appear in multiple positions and are not predictable for users, and all names with a variant STEM, as the STEM taxon cannot be determined if the STEM is not present. The number of INNS, and examples of each, are given in Table 4.11. As Appendix A11 shows, the number of INNs that are complex and irregular (types 3 and 4) are steady throughout the decades, whereas the number of type 2 (names with a STEM and a SUBSTEM) have been increasing.

Table 4.11 Number of INNs in each taxon type

Taxon type	***Number of INNs***	***%age***	***Example***
1 single-level taxon	151	33.4%	-tinib, for tyrosine kinase inhibitors
2 Multi-level	90	19.9%	-ase, for enzymes
3 Complex	62	13.7%	-gene, for substances for gene therapies
4 irregular	149	33.0%	fos, for insecticides, anthelminthics, pesticides etc.

4.4.3 STEMS used in INNs

The WHO Stem Book provides no explanations of the motivation of STEMS, i.e. why those particular forms were chosen to represent those pharmacological groups. Instead, the formal motivation can be observed and deduced. INNs are not full neoclassical compounds in that they cannot be broken down into elements that are derived from classical languages, but they *are* highly compositional and constituent parts can be analysed for their meanings (e.g. Deléger et al., 2009). INNs comprise:

- A "random, fantasy PREFIX" (WHO, 2017a), although in practice, this is not always at the beginning
- A STEM, typically at the end of the name
- A SUBSTEM, or multiple SUBSTEMS, typically in the middle of the name

Anecdotally there are some interesting backstories for the random, fantasy element (usually a PREFIX), such as carfilzomib (a multiple myeloma drug), named after molecular biologist Philip Whitcome and his wife, Carla, who both died from cancer (Drahl, 2012). Warfarin was named using an acronym of Wisconsin Alumni Research Foundation, a key funder in its development.

This chapter looks closely at the formal motivation of the 46 STEMS that occur in five or more INNs in this corpus, and some of their SUBSTEMS. These STEMS most commonly abbreviations or acronyms of longer terms. The mini-description of each STEM presented in the Stem Book (WHO, 2021b) was used to deduce the formation of the STEM based on its pharmacological group. For example, the title for monoclonal antibodies, with STEM -mab:

BAN, USAN

-mab **monoclonal antibodies** (see also Annex 3)

Many STEMS are ostensibly abbreviations of the pharmacological group, as described in the Stem Book, such as:

- vir for "antivirals";
- gest for "steroids, progestogens";
- gab for "gabamimetic agents";
- gli for "antihyperglycaemics" (with a vowel change from <y> to <i>, in line with prohibited letters);
- -cog for "blood coagulation factors" (with a vowel drop of the <a>)
- -stat for "enzyme inhibitors", presumably related to 'statins'

Some STEMS are partial acronyms of the pharmacological group. They are not standard acronyms using just the initial letters of words, but are combinations of letters pulled out of several words so as to form a pronounceable form of the STEM. Examples are:

- -mab for “monoclonal antibodies”;
- -rafenib for “Raf (rapidly accelerated fibrosarcoma) kinase inhibitors”
- -tinib for “tyrosine kinase inhibitors”
- -imus for “immunosuppressants (other than antineoplastics)”

Some STEMS are ostensibly abbreviations based on a chemical name, such as:

- sal for “salicylic acid derivatives”;
- -retin for “retinol derivatives”;
- -parin for “heparin derivatives including low molecular mass heparins”;
- -azosin for “antihypertensive substances, prazosin derivatives”;
- -cillin/-cillinam for “antibiotics, 6-aminopenicillanic acid derivatives” (with a vowel change from <a> to <i>);
- -calci- for “vitamin D analogues/derivatives”, such as tacalcitol and paricalcitol, presumably related to *calcium*.

Some STEMS derived from the chemical structure of the compound, such as:

- -ase for “enzymes”, as this is the standard SUFFIX STEM for any enzyme in chemical nomenclature, such as lipases and transferases.

Some STEMS are abbreviations of an INN, typically the first name of the STEM taxon published by the programme. This suggests that the STEM was backformed from the first name of the taxon, to produce the STEM, which was used in subsequent names in this taxon. Examples are:

- -ifene for “antiestrogens or estrogen [*sic*] receptor modulators, clomifene and tamoxifen derivatives” (clomifene was published in 1965, the first in the corpus with this STEM)
- -oxetine for “serotonin and/or norepinephrine reuptake inhibitors, fluoxetine derivatives” (fluoxetine was published in 1976, the first in the corpus with this STEM)
- -cycline for “antibiotics, protein-synthesis inhibitors, tetracycline derivatives” (tetracycline was published in 1959, the second in the corpus with this STEM, after oxytetracycline)
- -azepam for “diazepam derivatives” (diazepam was published in 1965, the first in the corpus with this STEM)

Some of the early INNs that prompted the formation of later STEMS have since been withdrawn from use, such as:

- ibufenac (in -ac for “anti-inflammatory agents, ibufenac derivatives”) - it was published in 1966, earlier than any other names in the corpus with this STEM, but was withdrawn for hepatotoxicity;
- piperazine (in -izine for “diphenylmethyl piperazine derivatives”) - it was published in 1953 but not as an INN, earlier than any other names in the corpus with this STEM, but is now restricted to animal use in the UK. Interestingly, piperazine was named for its chemical similarity to piperidine, part of piperine in the black pepper plant, with the INFIX of -az- denoting the extra nitrogen bond, since *azote* is an alternate term for nitrogen.

With these INNs, the reason for that particular STEM being formed is unclear. For many of the most frequently-used STEMS, their formal motivation is unclear, such as:

- -pril for “ACE inhibitors”, such as captopril, imidapril, and enalapril
- -terol for “bronchodilators, phenethylamine derivatives”, such as salmeterol and olodaterol

As outlined in previous work (Bryan et al., 2015), some STEMS, such as -mycin for antibiotics, are derived from the name of the bacterium genus. Some are partial neoclassical forms, drawing upon elements of chemical nomenclature (e.g. the STEM -fos, from Latin phosphorous, in clofenvinfos, denoting phosphorous derivatives) but adapting the orthography in line with WHO guidelines on prohibited letters. All of these are subjective judgements based on descriptions given in the Stem Book, written by the WHO INN programme. None of these STEM formation mechanisms are described by the WHO, they have only been deduced for this analysis.

4.5 Relational features

The relational measures in this analysis (Rel-01-Rel-10) pertain to how INNs in the dataset interact with each other. They are *not* related to WHO guidance or information in the Stem Book.

4.5.1 Selecting and measuring relational variables

Ten relational measures were explored. Initial and terminal bigrams and trigrams were explored, essentially the frequency of two- and three- letter strings at the beginnings and

endings of INNs (Rel-01, Rel-02, Rel-03, and Rel-04) to replicate the measures taken in a previous study by Lambert et al. (2001a). The frequency of the pharmacological STEM in the database (Rel-05) was then explored. To learn more about the cross-use of taxonomies such as INN and ATC (Anatomical, therapeutic, chemical), the alignment between INN classification system and the ATC taxonomy (Rel-06, Rel-07) was measured. Neighbourhood density and neighbourhood frequency are key similarity measures used in other studies of LASA drug names (e.g. Lambert, Chang, & Gupta, 2003; Lambert et al., 2001a). Finally, semantic transparency was explored using positional family size (Rel-10). Selection and coding of these variables is described in the following subsections.

4.5.1.1 Rel-1, Rel-02, Rel-03, and Rel-04, initial and terminal bigrams and trigrams

The frequency of the two- and three- letter strings at the beginnings and endings of INNs were counted in the dataset. Shared strings at the beginning and ending of names drive the risk of LASA error (Lambert et al., 1999; Shah et al., 2017; Thompson, 2008); thus, names with frequently used strings may be predictors of LASA error, since there are more names with which they can be confused. Beginnings and endings of names were spliced into two-letter strings (bigrams, i.e. two letter ngrams) and three-letter strings (trigrams, i.e. three-letter ngrams). These were ranked by frequency and proportion of total ngrams in that position, and the level of homogeneity was explored.

4.5.1.2 Rel-05, STEM frequency

This counts how many INNs use a STEM in the dataset, e.g. glecaprevir contains the STEM vir, which appears in 42 INNs. This is useful in inferring how familiar users will be with a STEM, and also useful in looking at how INNs within a STEM taxon have shared characteristics.

4.5.1.3 Rel-06, Rel-07, alignment between INN taxonomy and ATC classification index

To explore the structure of ATC (anatomical, therapeutic, chemical) codes for the INNs in the dataset, ATC family size ratio is calculated. As each name is presumed to have a one-to-one correspondence with its referent medication (synonymy or semantic drift is not present within a nomenclature), it can be assumed that the corresponding names for each medication can be represented in this hierarchy - i.e. each medicinal product has one name, and each name refers to a single medicinal product. Both the INN and the ATC systems can be used to parse (break down) the names according to the meanings of each of their constituent parts. The ATC family size ratio determines how typical a particular ATC classification is for this

STEM taxon, i.e. do most of the INNs in the taxon share this ATC classification or is it an anomaly?

The extent to which the ATC classification system is aligned with the STEM taxonomy displayed in INNs is calculated here by ATC family size. This variable describes the extent to which the ATC code is predictable from the INN STEM. Predictability is associated with a positive number. This is calculated as in Equation 4.1 and looks at higher-level ATC classification with just the top two levels (Rel-06), as well as the full ATC classification with all four levels (Rel-07). It is expected that ATC family size will be higher at the broader level than at the more specific four levels.

Equation 4.1 Calculation of ATC family size

$$\text{ATC family size} = log \frac{INNs\ sharing\ ATC\ code}{INNs\ not\ sharing\ ATC\ code}$$

Since the data are available, it is useful to look at similarity between the WHO and ATC classifications both broadly and granularly. For example, infliximab (L04AB02) can be broken down into four levels:

- L ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS
- L04 IMMUNOSUPPRESSANTS
- L04A IMMUNOSUPPRESSANTS
- L04AB Tumor [*sic*] necrosis factor alpha (TNF-alpha) inhibitors

As a member of the -mab taxon, infliximab (L04AB02) shares the top two levels of classification with 11 names, including upadacitinib (L04AA44) and ixekizumab (L04AC13). They all use L04 in their code. At the full four levels, infliximab (L04AB02) shares all four levels of ATC classification with two other INNs in the taxon: adalimumab (L04AB04) and golimumab (L04AB06). Under the broader ATC measure, infliximab scores -0.60, and under the specific ATC measure, it scores -1.18 - it is less predictable, as the taxon is more heterogeneous at this level. For names with more than one ATC code, scores were calculated for each ATC code and used to derive a mean.

4.5.1.4 Rel-08, Rel-09, neighbourhood measures

Because the study focusses on confusion between pairs of names, formal shared properties between pairs must be defined. Both orthographic (written) and phonological (spoken) similarity between pairs of names are accurate predictors of LASA errors (Lambert, 1997; Lambert et al., 2001b). The analysis of the nomenclature dataset therefore includes a

measurement of pairwise similarity, quantifying the level and extent of internal relational similarity in this system of nomenclature and exploring the features of names that are highly similar to each other, and those with a high level of distinction.

Neighbourhood density and neighbourhood frequency were key similarity measures used (Rel-08 and Rel-09). Single-name metrics of neighbourhood density and neighbourhood frequency are taken, so that standalone names can be characterised by the level of similarity they share with other names in the nomenclature. This replicates work from other studies (Lambert, Chang, & Gupta, 2003; Lambert et al., 2001a). A binary score was used that was close to the original method proposed by Andrews (1997). Previous work measured similarity using the Levenshtein edit distance (Bryan et al., 2015); this study has used the trigram-2b measure, which is predictive of error (Lambert et al., 1999). To complement neighbourhood frequency and to allow for the possibility that a positive neighbourhood frequency is due to a negligible difference in prescribing frequency, the binary score of 'top-prescribed neighbour' was also included, measuring whether any of the INN's neighbours are in the top 25 most prescribed medications in the dataset?

Pairwise similarity scores will be used to determine the neighbourhood of an INN, the density of that neighbourhood (how many similar names there are), and the frequency of the neighbourhood (whether the INN is prescribed more or less often than its neighbours). Typically a neighbour is defined as any name differing by a single letter (Andrews 1997) or an 'edit distance' of 1, but the operational definition is widened here to any trigram-2b score over 0.33, as in Lambert et al. (1999, 2004). Neighbourhood density and neighbourhood frequency are 3.1 and 3.2 in Table 4.12.

Trigram-2b is calculated by breaking two strings into three-letter sequences called trigrams and counting how many occur in both strings. Two spaces are added to the beginning of each string to add weight to the start, as names sharing initial letters have been found to be more likely to be confused (Lambert et al., 1999). For example, to compute the trigram-2b similarity between mercaptamine and mercaptopurine:

- mercaptamine (__m, _me, mer, erc, rca, cap, apt, pta, tam, ami, min, ine = 12 trigrams)
- mercaptopurine (__m, _me, mer, erc, rca, cap, apt, pto, top, opu, pur, uri, rin, ine = 14 trigrams)

The Dice coefficient (Equation 4.2) is used to compute orthographic similarity, expressed as the proportion of n-grams that are shared by the two strings (__m, _me, mer, erc, rca, cap,

apt, ine = 8 trigrams), where A is the number of trigrams in the first name, B is the number of trigrams in the second name and C is the number of trigrams that appear in both names. The score can range from 0 (no shared trigrams) to 1 (identical).

Equation 4.2 Trigram-2b calculation for mercaptamine and mercaptopurine

$$\text{Similarity} = \frac{2C}{A+B} = \frac{2*8}{12+14} = \frac{16}{26} = 0.61$$

Similarity was calculated using an Excel Macro, and scores were tabulated as in Table 5.3. Over half a million scores were calculated (between 1,046 names).

Table 4.12 Examples of pairwise scoring using the Trigram-2b measure

-	alphacetylmethadol	amfetamine	amfetaminil
alphacetylmethadol	100%	7%	7%
amfetamine	7%	100%	86%
amfetaminil	7%	86%	100%

These similarity measures are pairwise, i.e. they describe the similarity between two names. To transform this measure into a datapoint describing a stand-alone name, two measures were derived, as in Lambert, Chang, & Gupta (2003): neighbourhood density and neighbourhood frequency. Neighbourhood density is defined as the number of names with a certain level of orthographic similarity to each name, i.e. the size of its neighbourhood. The threshold for inclusion in the neighbourhood is a trigram-2b score of 0.33, as in Lambert, Chang, & Gupta (2003). Neighbourhood frequency is a binary measure of whether any of a name's neighbours is more frequently prescribed ('1') or not ('0'). It is generally accepted that words (extrapolated here to include drug names) with at least one higher frequency neighbour are harder to recognise than words with no higher frequency neighbours (Newbigging, 1961).

4.5.1.5 Rel-10, semantic transparency, measured by positional family size

Semantic transparency was measured using positional family size (PFS), which is a more specific measure of the predictability of a name's meaning. Although FREEFIXES are officially allowed within the INN nomenclature, if a STEM's position in the name varies this can cause confusion for users and make the name's meaning difficult to parse, e.g. *writer* and *writes* are easily recognised as a derivation and inflection of *write* because they appear in the same

place. PFS is a useful measure of this predictability, and is especially useful since it goes beyond the prescriptive guidance of the Stem Book, to look at actual presentation of INNs to users. When the head of a compound (the top category, e.g. BERRY in BLUEBERRY) is transparent and productive - i.e. it is the head of lots of compounds - this can reduce processing times, presumably because the user can more reliably infer its meaning (Feldman et al., 2004), i.e. because the STEM (and the head) being transparent facilitates meaning.

Semantic transparency has been found to be higher when components of a compound appear in their 'typical' position, as they are more productive in the nomenclature, and predictable for the user. Although this is in general-language English compounds, this finding might be transferable to neoclassical compounds, although they are different from other compounds in that they have fewer heads and more modifiers. Positional family size ratio (PFS ratio) was used to measure this predictability of meaning (Bell & Schäfer, 2016). PFS ratio is calculated as in Equation 4.3, which is adapted from the formula used in Schäfer (2018). A positive log indicates that the STEM is in its 'typical' position (1 was added to each side of the equation to account for any null values). Because some STEMS have more than one form, e.g. -azam in clobazam and -azolam in alprazolam, this measure included differing forms within a taxon, as well as position of the STEM, as a change in form would impede predictability as much as a change in position. The positional family size measure decouples semantics from the Stem Book, as it is based purely on the position of the STEM in the name, rather than comparing it to the prescribed position of the Stem Book. It is calculated as in:

Equation 4.3 Calculation of positional family size

$$log\ \frac{INNs\ with\ stems\ in\ this\ position}{INNs\ with\ stem\ in\ another\ position}$$

For example, lanreotide shows its STEM -tide at the end, as do all 17 of the INNs in the taxon, so its positional family size is positive, as in Equation 4.4. The position of the STEM is predictable. ‘1’ is added to each value to treat null values.

Equation 4.4 positional family size of lanreotide

$$log\frac{17+1}{0+1} = 1.26$$

On the other hand, the FREEFIX vir in efavirenz appears in the middle of the name, unlike most of the INNs in the taxon, and so its positional family size is negative, as in Equation 4.5 - the position of the STEM is not predictable.

Equation 4.5 positional family size of efavirenz

$$log\frac{8+1}{34+1} = -0.59$$

4.5.2 Distribution of relational variables

In this section a number of relational variables was explored within the dataset of English-language INNs. These features are listed in Table 4.13. Nine relational variables are continuous and measures of central tendency are given. For the one categorical variable, frequencies of each category are given, “yes” and “no”, excluding missing values.

Table 4.13 Summary of the distribution of relational variables in the taxonomic database

ID	Variable	Type of feature	Type of data	N valid cases [% of total cases]	Inclusion criteria	Descriptives / frequencies			
						Minimum	Maximum	Mean	Std deviation
Rel-01	How frequently do the first two letters of the name (initial bigram) appear in the data set?	Relational	Continuous	1,046 [100.0%]	All names	1.00	34.00	12.96	8.22
Rel-02	How frequently do the first three letters of the name (initial trigram) appear in the data set?	Relational	Continuous	1,046 [100.0%]		1.00	18.00	3.75	3.96
Rel-03	How frequently do the last two letters of the name (terminal bigram) appear in the data set?	Relational	Continuous	1,046 [100.0%]		1.00	312.00	133.63	121.35
Rel-04	How frequently do the last three letters of the name (terminal trigram) appear in the data set?	Relational	Continuous	1,046 [100.0%]		1.00	228.00	72.21	85.81
Rel-05	How frequently does this INN's STEM appear in the dataset?	Relational	Continuous	774 [74.0%]	Names that are listed as 'compliant' in the Stem Book	0.00	63.00	10.00	16.25
Rel-06	To what extent do the ATC codes assigned to the INN align with the ATC codes for other names in the same taxon? (top two levels of ATC)	Relational	Continuous	452 [43.2%]	Names that are listed as 'compliant' in the Stem Book and belong to a pharmacological	-1.50	1.61	0.04	0.49

Rel-07	To what extent do the ATC codes assigned to the INN align with the ATC codes for other names in the same taxon? (all four levels of ATC)	Relational	Continuous	452 [43.2%]	STEM taxon with at least five names	-1.50	1.08	-0.24	0.41
Rel-08	Neighbourhood density: how many names are similar to the INN?	Relational	Continuous	1,046 [100.0%]	All names	0.00	25.00	4.35	4.35
Rel-10	Semantic transparency, measured by positional family size ratio: to what extent is the INN's STEM in a typical position for names in the same taxon?	Relational	Continuous	452 [43.2%]	Names that are listed as 'compliant' in the Stem Book and belong to a pharmacological STEM taxon with at least five names	-0.81	1.81	0.35	0.58
						Yes [% valid cases]		No [% valid cases]	
Rel-09	Neighbourhood frequency: are any of the INN's neighbours prescribed more frequently?	Relational	Categorical, nominal [yes/no]	1,000 [95.6%]	Names with a prescribing frequency value	441 [44.1%]		559 [55/9%]	

4.5.2.1 Rel-1, Rel-02, Rel-03, and Rel-04, initial and terminal bigrams and trigrams

The most frequent initial trigram in this INN database is *flu-*, as in flumazenil (in 18 names, 1.7%), and the most frequent terminal trigram is *-ine* as in mercaptamine (in 228 INNs, 21.8%). Bigrams and trigrams were analysed in an earlier study by Lambert et al. (2001a) and measures in this database are compared with results in that study.

The most common initial bigrams and trigrams are given in Table 4.14, including comparable ngrams identified by Lambert et al. (2001a). Figures 4.15 and 4.16 illustrate the distribution of initial bigrams and trigrams. Most common is defined as in the 'top ten', although some include a 'top twelve' when some strings appear in equal number (for example, in Table 4.14 al-, be-, and le- are initial bigrams that appear in equal number).

Table 4.14 Top ten beginnings of INNs, compared with those of USANs (initial bigrams and trigrams)

This analysis (n = 1,046)						**Lambert et al. 2001 (n = 8,712)**					
Bigram	***n***	***%***	***Trigram***	***n***	***%***	***Bigram***	***n***	***%***	***Trigram***	***n***	***%***
me-	34	3.3	flu-	18	1.7	me-	228	4.28	met-	107	2.01
ca-	31	3.0	pro-	16	1.5	di-	154	2.89	sul-	90	1.69
pr-	28	2.7	car-	14	1.3	pr-	126	2.36	flu-	90	1.69
tr-	24	2.3	met-	14	1.3	su-	121	2.27	clo-	84	1.58
te-	23	2.2	chl-	12	1.1	ni-	121	2.27	pro-	74	1.39
di-	20	1.9	clo-	12	1.1	tr-	117	2.19	tri-	68	1.28
le-	20	1.9	tri-	11	1.1	cl-	115	2.16	fen-	54	1.16
be-	19	1.8	cef-	10	1.0	pi-	115	2.16	ben-	62	1.07
da-	19	1.8	lev-	10	1.0	fl-	115	2.16	car-	57	1.01
fl-	19	1.8	hyd-	9	0.9	be-	110	2.06	cef-	49	0.92
mi-	19	1.8	ben-	8	0.8						
			tra-	8	0.8						
TOTAL	256	24.47		142	13.59		1322	15.17		735	8.44

Figure 4.15 Density plot for Rel-01, initial bigrams

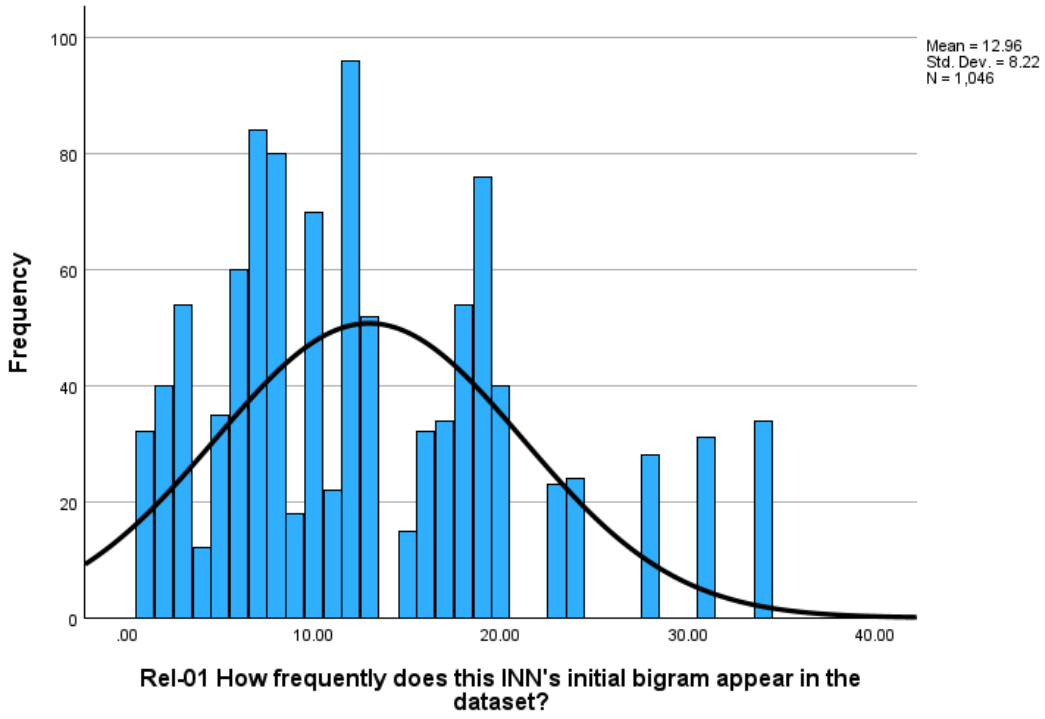
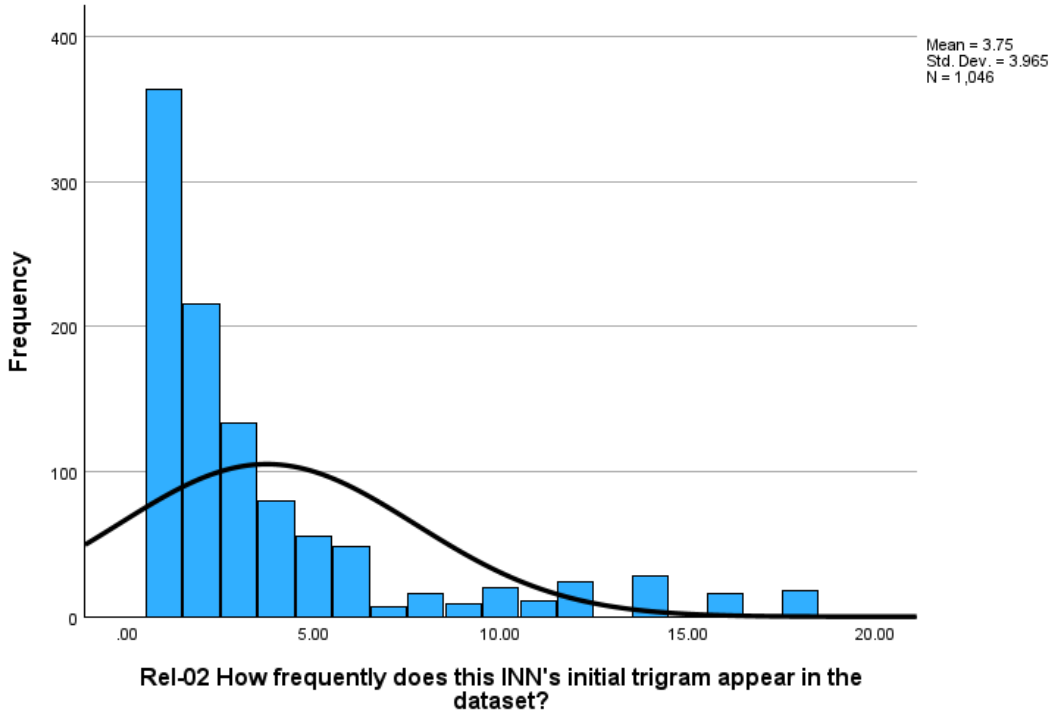


Figure 4.16 Density plot for Rel-02, initial trigrams



There is much less heterogeneity in the terminal strings than in the initial strings, with over three quarters of all INNs ending in one of the top ten bigrams (e.g. mercaptamine ending in -ne) and over half ending in one of the top ten trigrams (e.g. aciclovir ending in -vir). Table

4.15 compares terminal bigrams and trigrams between the current INN database and the database of USANs in Lambert et al.'s 2001 study (2001a). Figures 4.17 and 4.18 illustrate the distribution of terminal bigrams and trigrams In some cases this is unavoidable, as -ne forms part of the name of the chemical compound, such as -purine and -amine. Initial strings show variation: the most common initial bigrams make up a quarter of INNs (e.g. diazepam beginning in di-), and the top ten initial trigrams make up an eighth (e.g. flumazenil beginning in flu-). The most common ending is -ine, appearing in 21.8% of INNs, such as mercaptopurine. Coverage of the top ten initial bigrams and trigrams is lower in Lambert's 2001 analysis of USANs than in the present analysis of INNs, pointing to more variation in the USAN nomenclature than in the INN nomenclature.

Table 4.15 Top ten endings of INNs, compared with those of USANs (terminal bigrams and trigrams)

This analysis (n = 1,046)						**Lambert et al. 2001 (n = 8,712)**					
Bigram	***n***	***%***	***Trigram***	***n***	***%***	***Bigram***	***n***	***%***	***Trigram***	***n***	***%***
-ne	312	29.8	-ine	228	21.8	-ne	1844	34.59	-ine	1191	22.34
-in	128	12.2	-ide	104	7.9	-in	626	11.74	-ide	553	10.37
-de	83	7.9	-one	64	6.1	-de	560	10.50	-one	504	9.45
-ol	77	7.4	-mab	63	6.0	-ol	478	8.97	-ate	234	4.39
-ab	63	6.0	-nib	38	3.6	-te	243	4.56	-ole	222	4.16
-ib	52	5.0	-cin	36	3.4	-le	227	4.26	-cin	198	3.71
-an	35	3.3	-vir	33	3.2	-il	163	3.06	-nol	110	2.06
-il	33	3,2	-ole	30	2.9	-am	128	2.40	-lin	94	1.76
-ir	33	3,2	-tin	22	2.1	-an	120	2.25	-ene	89	1.67
-le	30	2.9	-tan	19	1.8	-en	117	2.19	-lol	85	1.59
TOTAL	846	80.88		637	60.90		4506	51.72		3280	37.65

Figure 4.17 Density plot for Rel-03, terminal bigram

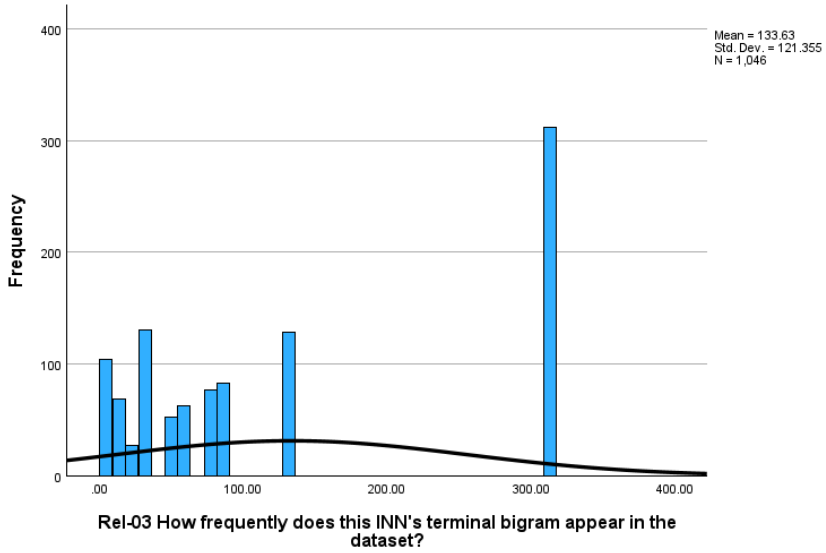
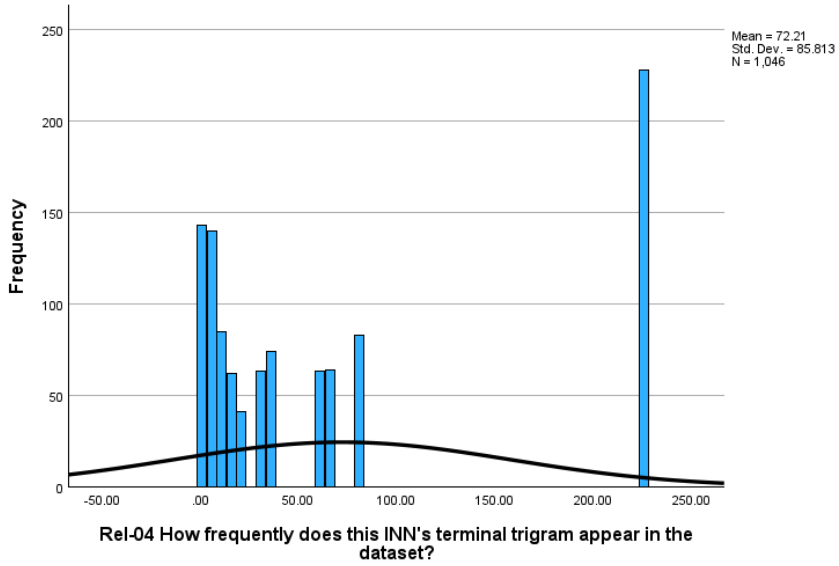


Figure 4.18 Density plot for Rel-04, terminal trigram

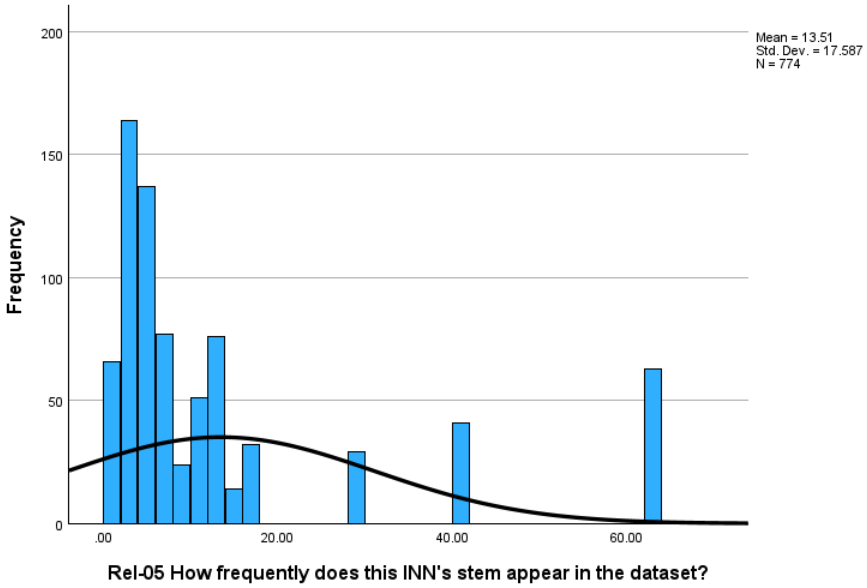


4.5.2.2 Rel-05, STEM frequency

A density plot for STEM frequency is given in Figure 4.19. After manually parsing each name (looking it up in the Stem Book and identifying its STEM [and any SUBSTEMS]), 197 unique STEM taxa are identified, comprising multiple SUBSTEMS. Monoclonal antibodies (STEM -mab) are the most frequent, in 62 of 1,162 INNs (5.94%), followed by antivirals (STEM vir), in 42 INNs (3.61%) and antineoplastics (STEM -tinib), in 29 INNs (2.51%). Following the methodology of Schäfer (2018), only those STEMS used in at least five INNs are included in some aspects of the analysis, namely the positional family size and ATC family size - this

leaves 41 STEMS, used in a total of 452 unique INNs, 43.2% of the corpus. Names were marked as belonging to a frequent STEM taxon, or not. STEM frequency of INNs has been increasing since the 1980s (see Appendix A11), likely due to the large number of monoclonal antibodies (-mab), which have since been split into four separate STEM taxa.

Figure 4.19 Density plot for Rel-05, STEM frequency



4.5.2.3 Rel-06, Rel-07, alignment between INN taxonomy and ATC classification index

The extent to which the ATC codes of names in a STEM taxon are homogeneous was also measured, to the top two levels of ATC (a broad measure signalling broad alignment, Rel-06) and to all four levels of ATC (a granular measure signalling full alignment, Rel-07). Some medications are not assigned an ATC code[14] (e.g. carbidopa), whereas others may be assigned multiple ATC codes (e.g. neomycin with S01AA03, A07AA01, A01AB08, B05CA09, D06AX04, J01GB05, R02AB01, S02AA07, and S03AA01, because it has numerous formulations and applications). To look at the extent to which INNs within a taxon share ATC codes, only the taxa are selected that had both at least 5 INNs with that STEM, and at least 5 ATC codes assigned to those names. There are 41 STEMS that have at least five INNs, with at least five ATC codes. As shown in Appendix section A12, there is variation in the extent to which ATC codes are shared throughout each taxon. Some taxa demonstrate high degrees

[14] 10 names: cocarboxylase, crisantaspase, fosaprepitant, gimeracil, netupitant, oteracil, mestranol, poloxamer, carbidopa, carbomer

of similarity in that they share multiple levels of ATC codes (e.g. -ac and -pril in the figure) - others show low similarity in sharing only the upper levels of ATC codes (e.g. -dil in the figure). Some show a balance of the two, such as -mab in the figure. For the larger taxa with more ATC codes, the number of shared ATC codes reduces down the levels of ATC system - names tend to share the top level (anatomic target) but they then start to branch out into various therapeutic functions and chemical attributes.

Broad and specific alignment was measured. Broad alignment was measured by looking at whether the ATC top two levels of the INN are typical for names in the STEM taxon, by taking a family size ratio score. Scores less than 0 indicate the ATC codes (top two levels only) are not typical, and scores more than 0 indicate they are typical. Specific alignment was measured in the same way, but taking into account all four nested levels of the ATC taxonomy. Mean score for all INNs is 0.10 (SD 0.74) for broad alignment, and -0.57 (SD 0.46) for specific alignment. Broad alignment will be higher because several ATC codes within a STEM taxon may share top level codes (such as all being cardiac therapies, code C01).

Figure 4.20 Density plot for Rel-06, broad alignment with ATC index

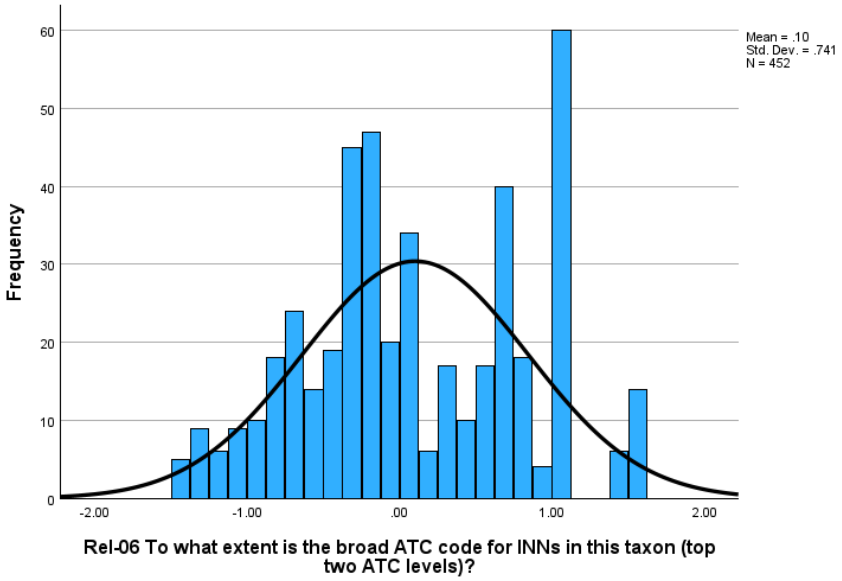
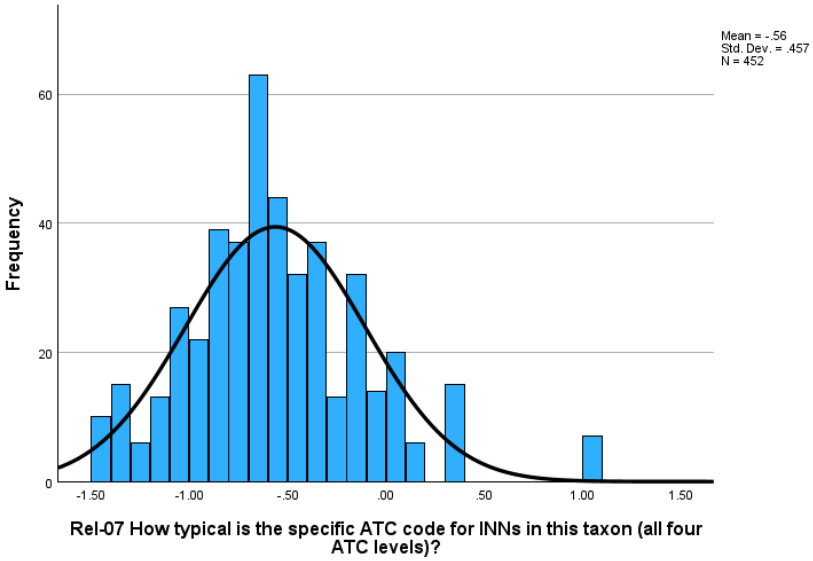


Figure 4.21 Density plot for Rel-07, specific alignment with ATC index



In the early decades of the INN programme, the broad alignment of names with the ATC taxonomy was lower, but has remained steady since the 1970s. However, the specific alignment has dropped in recent decades, suggesting increased heterogeneity in ATC third and fourth level codes (see Appendix A11).

4.5.2.4 Rel-08, Rel-09, Rel-11, neighbourhood measures

A neighbourhood is defined here as those names that have a similarity score (in this case, using the trigram-2b measure) of 0.33 to a particular name. The neighbourhood density (ND), therefore, is the number of names in the neighbourhood. Of 1,046 INNs in the full dataset, 111 have an ND of 0, and a further 610 names have an ND of between 1 and 5. The modal INN has one neighbour (i.e. a neighbourhood density of 1), and one name, mepyramine, has 30 neighbours. Figure 4.22 gives a breakdown and Figure 4.23 gives a density plot. Names were binned according to their neighbourhood density and explored the STEMS used in names with a high number of neighbours (n=42, those with at least 16 neighbours) and no neighbours (n=115). Names with a high number of neighbours overwhelmingly use either the -mab (21 of 46) or -tinib (13 of 46) STEM, which are the most commonly used STEMS in the dataset (examples: natalizumab, palivizumab, reslizumab). Names with no neighbour at all feature a total of 52 STEMS, including montelukast (-ast) and pegaspargase (-ase), but also a large proportion of names with no STEM taxon (43).

Figure 4.22 Density plot for neighbourhood density

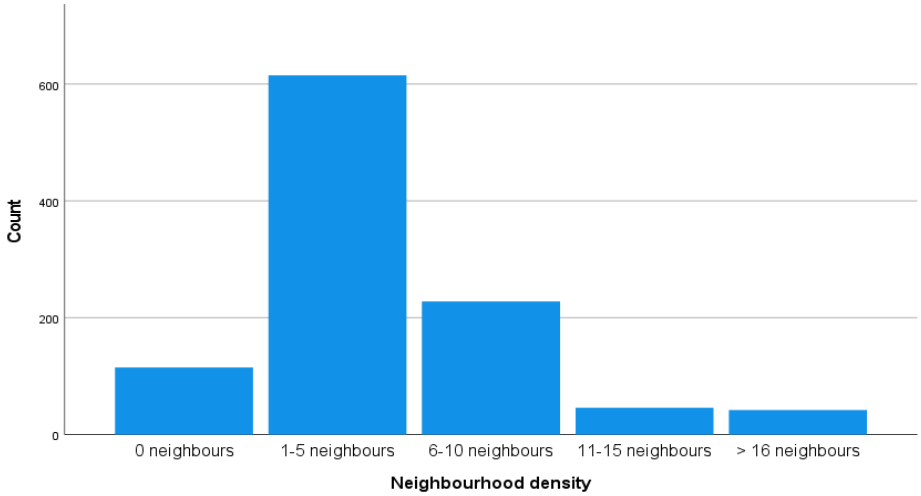
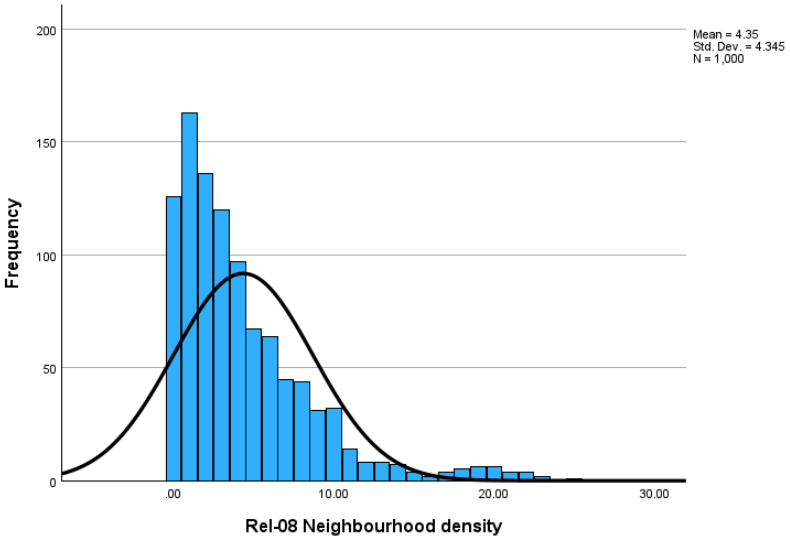
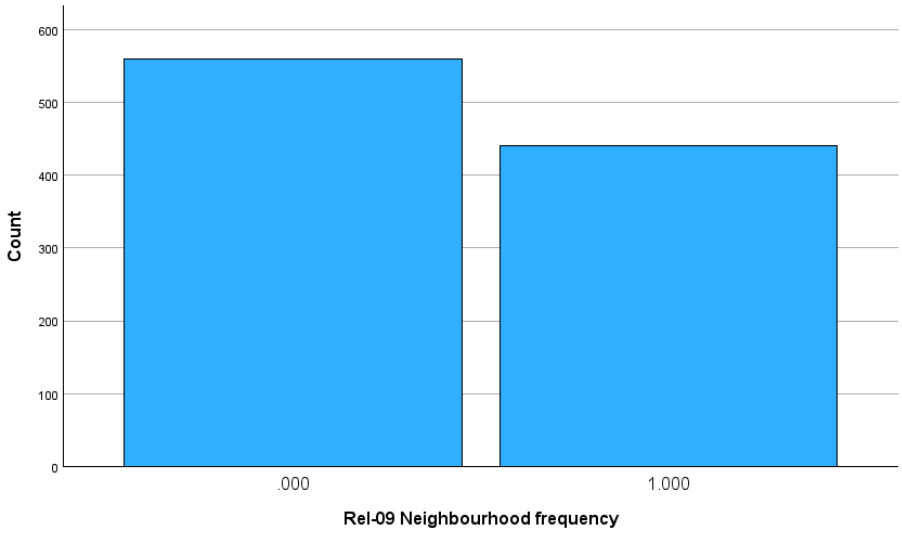


Figure 4.23 Density plot for Rel-08, neighbourhood density



Neighbourhood frequency was also measured. This is a binary measure used by Andrews (1997) and Lambert, Chang, & Gupta (2003), and the operational definition is to assign a neighbourhood frequency value of '1' if at least one neighbour has a higher frequency, and in other cases to assign a neighbourhood frequency value of 0 (Andrews, 1997). Of the total dataset of INNs, 559 (53.4%) have a neighbour with a higher prescribing frequency (score is 1) and 421 (42.2%) have the highest prescribing frequency of their respective neighbourhoods (score is 0). For example, metformin has a log prescribing frequency of 7.68; the four names in its orthographic neighbourhood have the prescribing frequencies: mecasermin (3.62); meglumine (7.38); metaraminol (6.48); and methenamine (5.46). 113 (10.8%) names have a neighbour that is 'super-prescribed', i.e. names of the top 25 most-prescribed drugs in the UK, e.g. omeprazole. (this is Rel-11).

Figure 4.24 Bar chart for Rel-09, neighbourhood frequency

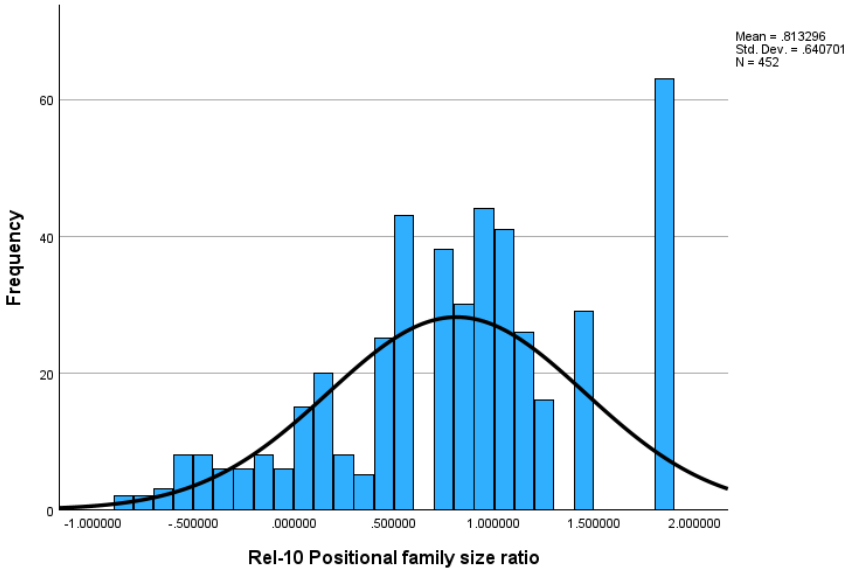


4.5.2.5 Rel-10, semantic transparency, measured by positional family size

Positional family size[15] is a useful measure of the predictability of a name's meaning, and is furthermore useful since it goes beyond the prescriptive guidance of the Stem Book, to look at the actual presentation of INNs to users. Following the methodology of Bell & Schäfer (2016), only frequent STEMS (with at least 5 names) were included in the measure. Positional family size ratio is calculated for 452 unique names, as in Equation 4.3 in the previous section. The dataset as a whole has a mean positional family size of 0.81, SD 0.64 (a value of more than 0 means the STEM is in a typical, thus predictable position). This means that the semantic head of INNs is largely predictable, with STEMS generally following a predictable positioning for the user. Average predictability has steadily increased since the 1970s, as in Appendix A11.

[15] Measure Rel-10

Figure 4.25 Density plot for Rel-10, semantic transparency



4.6 Discussion, conclusion, and limitations

The results from analysis of each measure are presented in the previous sections; here a summary of results is provided, and core findings are presented. Analysis was undertaken in three parts: orthographic measures, morphosemantic measures, and relational measures.

4.5.1 Descriptive vs prescriptive

Terminology and nomenclatures should work onomasiologically, seeking to attribute a term to an existing concept in a systematic and predictable way (Cabre 1999: 38). Many of the INNs in the corpus are absent from the Stem Book (256/1,046 names, or 24.47%), and those remaining have a high level of adherence to Stem Book guidance (Morph-01, only 80/916 names, or 8.73%, do not adhere to guidance). It is impossible to determine whether those absent adhere to guidance or not, but previous work found inconsistent adherence to naming guidelines (Bryan et al., 2015). Although the ordering of STEMS within names varies, and headedness is both left and right, adherence to the guidance in the Stem Book is high. It may be that the high level of adherence to positioning, headedness, and use of STEMS in the guidance is because the Stem Book serves a *de facto* descriptive function, rather than prescriptive function; i.e. that the Stem Book describes the existing nomenclature, rather than dictates the formation of future nomenclature. This begs the question: where is the guidance for using STEMS in INNs?

4.5.2 Semantic transparency vs minimising similarity

Previous work in this area identified a tension between using STEMS to increase users' understanding and minimising orthographic (and phonological) similarity (Bryan et al., 2015). Several results in this chapter strengthen this observation. For example, looking at ngram frequencies (Rel-01 – Rel-04), the majority of INNs share a common (in the top ten) ending, both two- and three- letters, e.g. -ne or -vir respectively, but they do not share a common beginning. Looking at Morph-05, most INNs incidentally contain another STEM, but do not form part of that pharmacological group. In part, the homogeneity in endings is due to names employing chemical terms, such as -amine and -purine. More broadly, however, there is much more heterogeneity in the beginnings of INNs, because most STEMS appear at the end - most INNs are right-headed (Morph-07). Looking at neighbourhood density (Rel-08), the majority of INNs have at least one neighbour - and many have more than ten neighbours - so orthographic similarity is widespread through the corpus.

The results from several measures indicate that semantic transparency is high in many INNs, in particular STEM ordering (Morph-02, Morph-03), headedness (Morph-07), FREEFIXES (Morph-06), and positional family size (Rel-10). Most INNs have a head (a hyperonym, or 'parent category' STEM) – only 3.71% of INNs in the sample have no hyperonym STEM (such as pivmecillinam, with SUBSTEM -cillinam, which has no 'parent' STEM). In those with a hyperonym STEM, a tendency for right-headedness in the nomenclature reflects typical right-headedness in English general language, which will increase predictability of meaning for users (they interpret the ending as the 'top category', and then move to the left to sub-categorise. This contributes to increasing semantic transparency - the predictability of meaning based on the form. The low prevalence of FREEFIXES is also encouraging (only 18.42% - or 154/836 INNs - of the sample contains a FREEFIX STEM).

The prime measure in this analysis for semantic transparency is positional family size (Rel-10), and the mean score here is positive. This indicates that although not all INNs are right-headed, there is internal consistency within STEM taxa. For example, within a STEM taxon, if the STEM appears at the beginning of the INN, typically all INNs within that taxon will be left-headed. This suggests that if the user were to be aware of the 'official' position of the STEM, this would increase the predictability of meaning for names - but this adds to the burden of learning on the user and presumes some level of knowledge of word formation. Modal ATC family size (Rel-06, Rel-07) was negative, which points to a disconnect between the ATC and INN classification systems. If users are familiar with the ATC taxonomy, alignment with the INN taxonomy will increase their understanding of STEMS and potentially increase

semantic transparency, on a user level. Reduced ATC family size (i.e. INNs in a STEM taxon do not typically share ATC codes) will decrease semantic transparency but it is difficult to quantify the extent of this effect, and the extent to which this counteracts the positive effects of other linguistic features.

Thus, there is a tension between promoting semantic transparency (maximising predictability through using STEMS and maintaining right-headedness) and minimising similarity that leads to confusability (by promoting distinct orthography). This was initially observed in earlier work (Bryan et al., 2015), but these results give a more detailed and granular analysis.

4.5.3 Conclusion and limitations

This chapter provides descriptive statistics from the analysis of 22 linguistic variables. The statistical significance of these results, and their relationship with the likelihood of error, will be explored and discussed in Chapter 5. Measuring the number of syllables was complicated by a number of double or triple vowel combinations, some of which are pronounced as a single syllables (e.g. parathyrOId) and some of which are sometimes, but not always, pronounced as multiple syllables (e.g. phenytOIn). Since there is no central or official guidance on pronunciation of names, the syllable measure of these vowel clusters was decided subjectively and gives room for error, disagreement, and allowance should be made for regional variants. The researcher making these subjective judgements speaks British English, with approximate Received Pronunciation, so pronunciation may conflict with US, Canadian, or Australian English, and with the global English spoken by millions as a second language around the world. The semantic parsing of names was wholly reliant on the Stem Book, but the guide itself is confusing in parts and does not contain all INNs, so it is of limited reliability.

Chapter 5: Predicting the occurrence of LASA errors in English

5.1 Introduction

This thesis builds knowledge of the linguistic characteristics of generic drug nomenclature and explores the relationship between these characteristics and LASA errors, thus refocussing on drug naming as a key stage in error mitigation. Chapter 4 characterises the data, describing linguistic features of a database of International Nonproprietary Names (INNs). This chapter explores the association between those linguistic features and LASA errors, working to pinpoint linguistic features of INNs that act as latent conditions in the system (Reason, 2000) and increase the probability of Look-Alike Sound-Alike (LASA) errors. This chapter addresses Research Question 2 (RQ2): *To what extent are any of the linguistic features described in answer to RQ1 correlated with the occurrence of LASA errors?*

Previous chapters and published work have reviewed the literature on LASA errors and their potential solutions, both narratively and systematically (Bryan et al., 2015, 2020b, 2020a). These chapters and publications identified a lack of focus in the literature on the drug naming stage and an undue shift in responsibility towards healthcare practitioners and away from manufacturers and naming bodies such as the World Health Organization (WHO). Chapter 4 presented several interesting findings. Firstly, although headedness and positioning of names varies, this variance is worked into the WHO naming guidelines and so adherence to guidance is high. This suggests that guidance may be descriptive rather than prescriptive, because the naming guidelines are incrementally updated but describe names that are up to 75 years old. Secondly, although the endings of INNs are highly homogeneous (132 unique terminal trigrams) the beginnings are more diverse (567 unique initial trigrams), similar to that observed in Lambert et al. (2001a). thirdly, most names have at least one orthographic neighbour (a name that is highly similar to it in its written form) and some have more than 20 orthographic neighbours. Fourthly, there is low alignment between the INN and ATC classification systems. Finally, semantic transparency appears high in the dataset, with a high average positional family size ratio of 0.83 ±0.64, where values over 0 imply semantic transparency (all values range between -1 and 1).

Based on the findings of the relevant literature and the previous chapter, there are some tentative *a posteriori* assumptions of features that will be associated with LASA errors. Firstly, names containing high-frequency initial strings (initial bigrams and trigrams), as found by Shah et al. (2017). Secondly, names with a high number of orthographic neighbours (a so-called *high neighbourhood density*), because orthographic similarity increases the likelihood of LASA errors. This was found in the US nomenclature by Lambert et al. (2001a). Thirdly, names with a neighbour that is prescribed more frequently (so-called *positive neighbourhood frequency*), as this has been shown to increase the likelihood of LASA errors (Lambert, Chang, & Gupta, 2003). Fourthly, names that are semantically transparent, because semantic transparency has been found to reduce the time needed to correctly process a compound (Bell & Schäfer, 2016; Libben et al., 2003), though there is no experimental research on semantic transparency in INNs or on how semantic transparency affects differentiation between two similar words or terms (see section 4.5.1.5 for a detailed discussion).

5.2 Methodology

The goal of this statistical analysis is to examine the relationship between linguistic features of INNs and the occurrence of LASA errors. To do so, an established analytical approach is adopted, following a series of recommendations by Heinze et al. (2018). Briefly, this comprises assessing variables for collinearity and redundancy, fitting a model using backward elimination, then assessing the stability of that model using bootstrap resampling.

The outcome variable – whether a name has been reportedly involved in a LASA error – is binary (“yes” or “no”), so binary logistic regression is the most appropriate tool. The analysis is not intended to build a predictive model but rather to explore and describe the relationships between linguistic features and error occurrence, helping to identify potential risk factors. As Heinze et al. (2018: section 2.1.1) note, explanatory models are used to “estimate causal effects of risk factors or exposures by means of adjusted effect estimation”. The summary analysis and most of the statistical testing is performed in SPSS. The model building is conducted in R due to the availability of the necessary functions.

Since there are 23 independent variables in the taxonomic database described in Chapter 4, both categorical and continuous, and since there is no strong prior theoretical basis to favour some over others, variable selection is necessary to identify the most relevant predictors. However, all model selection methods come with challenges: they can overstate the importance of some variables, exclude others incorrectly, and introduce bias. Among these

methods, backward elimination (BE) is preferred over forward selection (Heinze et al., 2018; Mantel, 1970) because it begins with all the variables included, reducing the risk of omitting important ones too early. Given the exploratory nature of this analysis, BE provides a systematic way to assess which variables show meaningful associations with LASA errors. Building such models is challenging, as simple methods such as stepwise selection can be biased. On the other hand, complex methods such as LASSO (least absolute shrinkage and selection operator) can exclude relevant predictors (Hastie et al., 2016).

As with any dataset, there is a risk that the observed sample may not perfectly represent the broader population of interest, i.e. the full nomenclature of INNs, and more broadly generic drug names. To address this risk, the stability of the final model is assessed through bootstrap resampling and related methods. Resampling creates multiple alternative datasets by drawing repeatedly from the observed data, allowing an evaluation of how often key variables occur across different iterations. If the same variables remain consistently selected, this suggests the model is robust to variations in the dataset. As Heinze et al. (2018, section 2.4.2) explain, model stability is a critical but often overlooked issue in variable selection: “bootstrap resampling with replacement or subsampling without replacement are valuable tools to investigate and quantify stability of selected models” (see also Sauerbrei et al., 2015).

The following subsections describe the process of preparing the dataset for analysis, including identifying and controlling for non-linguistic confounders, handling collinearity, and address missing data. The dataset will be structured to minimise collinearity, ensuring the best possible foundation for a robust explanatory model.

5.2.1 Preparing the dataset for analysis

This section presents the steps taken to structure and streamline the dataset for the statistical analysis. Firstly, non-linguistic confounders need to be accounted for in regression because they likely play a role in driving error. Secondly, some linguistic variables contain data that is not missing at random, so this must be accounted for in the regression while also reducing collinearity. Thirdly, many of the 23 linguistic variables presented in Chapter 4 are intrinsically correlated or related and pose a challenge to model development and interpretation, so a decision was made to remove these to avoid obvious duplication and the resultant collinearity. Variable names are given in a shorthand in this chapter, for brevity, e.g. $\text{Rel-02}_{\text{initial trigrams}}$.

A number of cases were excluded due to erroneous prescribing frequency values. Thus, the statistical analysis treats a subset of the dataset described in Chapter 4. This leaves exactly 1,000 names in the dataset for analysis. The outcome measure used, *error*, is binary: whether or not a name has been reported in a LASA error with another INN. 19.5% of names have been involved in a LASA error (195/1000).

5.2.1.1 Non-linguistic confounders

Several non-linguistic confounders were considered for inclusion in the analysis. These are not linguistic features but they are very likely to be associated with differences in LASA error rates. These confounders were identified as potentially relevant: prescribing frequency; the year the name was approved; and whether or not the name has been assigned an ATC (anatomical theoretical chemical) code. Prescribing frequency of names can be considered a proxy measure of lexical frequency, i.e. the usage of the names, since the more a name is prescribed, the more the name is spoken, written, heard, and read. Mathematically, the more times a medication is used, the more likely it is to have been subject to an error. The year in which the name was approved may also be relevant, since the longer a name has been in circulation, the more opportunities there have been historically for LASA errors to occur. Behaviour around prescribing may also change over time, and reporting differences across such a long timescale may introduce bias. Whether or not a name has an ATC code is relevant to analysis, since medicinal products that are used more frequently will be assigned ATC codes. Names in the dataset were approved from between 1955 and 2019 and the average prescribing frequency in the dataset is >6 million items prescribed per year in England and Wales. The majority of names in the dataset have an ATC code: 98.6% (986/1,000).

To ensure the linguistic variables in model selection are not affected by the presence of non-linguistic confounders, it is important to confirm these are confounders and then to determine how best to incorporate them into the model. Binary logistic regression was used to confirm their association with the outcome LASA error, both pairwise and with the three confounders combined, and results are given in Table 5.1. Prescribing frequency was treated as continuous data, natural numbers were not normally distributed (skewness 21.71, SE 0.08) and when transformed to their logs, skewness was reduced to -1.16 (SE 0.08). Year was treated as continuous data; and ATC code was treated as a binary factor (“yes” or “no”).

Prescribing frequency is significantly associated with error (p<0.001, OR 1.701 [1.468, 1.970]), with high prescribing frequency associated with a higher LASA error rate. The year the name was approved is also significant (p<0.001, OR 0.963 [0.952, 0.974]), with older names associated with a higher error rate. They will be included in the dataset for statistical analysis as $X01_{\text{prescribing frequency}}$ and $X02_{\text{year}}$ respectively. Whether or not the name has an ATC code is significant in isolation but drops out in the combined model, presumably because having an ATC code is strongly influenced by the age and uses of the drug.

It is not surprising that prescribing frequency (see Figure 5.1) is significantly correlated with the risk of LASA errors, since the more the name is used, the more opportunities will have arisen for it to be involved in an error. This also applies to year (Figure 5.2): newer names are less likely to be reported in a LASA error. This remains significant in the presence of prescribing frequency, suggesting that age is an independent factor in addition to prescribing frequency and may be tied to users' familiarity with the name. Names that were approved longer ago and have been in circulation for longer are prescribed more often. However, the relationship between prescribing frequency and year is not linear (Figure 5.3). Names approved in the most recent two decades (post-2000) are prescribed less often than names that were approved pre-2000, most probably because other drugs were available in generic formulations, and therefore at a lower cost to the NHS. Lower prescribing frequency post-2000 means that the risk of error is reduced in post-2000 names, with significant differences in error frequencies between names approved in the 1950s and names approved post-2000 (1950s vs 2000s, p=0.013, OR 0.323 [0.133, 0.785].

Table 5.1 Pairwise and combined model of four confounders and the outcome measure, using Binary Logistic Regression

	Pairwise		**With other two confounders**	
Variable	**p value**	**OR [CI 95%]**	**p value**	**OR [CI 95%]**
Prescribing frequency	<0.001	1.852 [1.600, 2.143]	<0.001	1.689 [1.456, 1.959]
Year	<0.001	0.965 [0.956, 0.974]	<0.001	0.962 [0.951, 0.973]
ATC code assigned	0.041	0.292 [0.090, 0.952]	0.893	1.093 [0.296, 4.040]

Figure 5.1 Prescribing frequency and LASA error

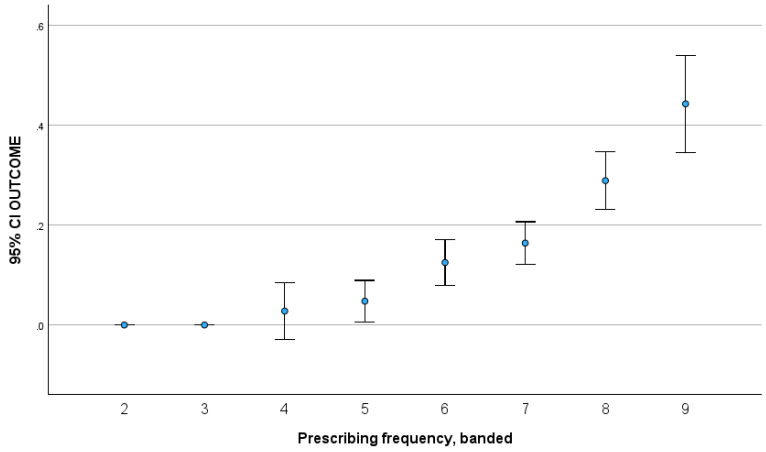


Figure 5.2 Decade and LASA error

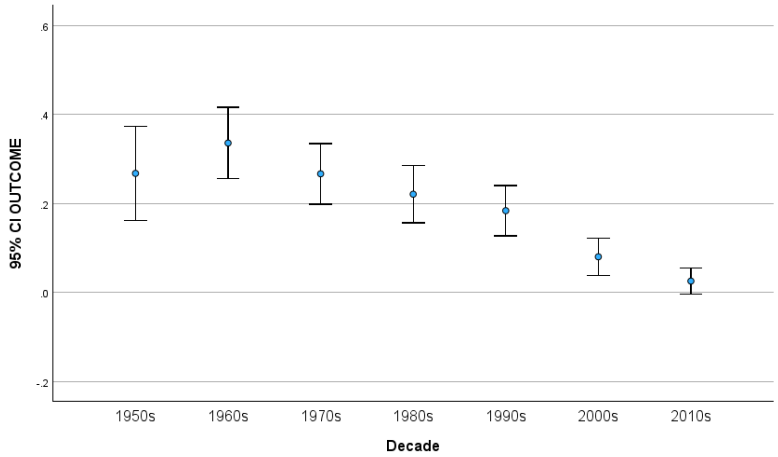
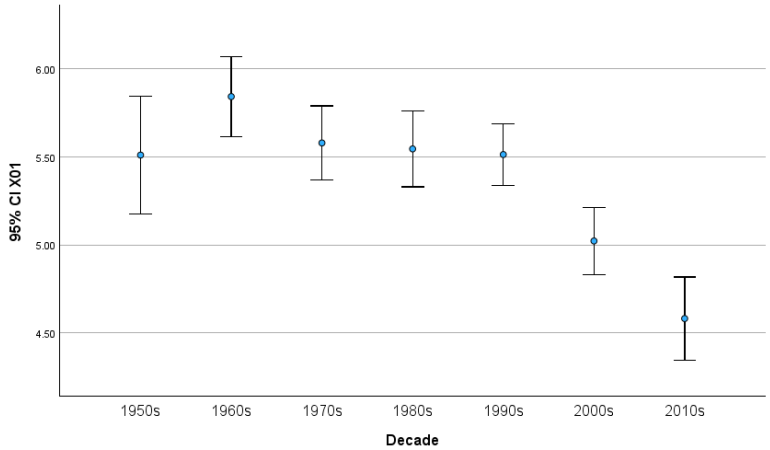


Figure 5.3 Prescribing frequency and decade



The non-linguistic confounders are then tested in a simple regression model with the error outcome to determine whether to treat them as linear covariates or factors. The process of assessing how to treat the confounders is outlined in Table 5.2. The contribution to risk of the confounders is of interest in the analysis, but the primary focus is on linguistic features – the confounders are in the model to remove their influence on the risk so that what remains is driven by the linguistic factors. Therefore, they need to be modelled in the most simple and accurate way, so their impact is accounted for but they are not overfitted. Based on AIC, prescribing frequency performs best as a log of natural numbers, treated as a linear covariate. It appears linear on a graph, as in figure 5.1. Year also performs best as a linear covariate (see figure 5.2). Prescribing frequency was log transformed from its natural numbers.

Table 5.2 Building the base model of three confounders (best performing highlighted in green)

Model	**N**	**Degrees of freedom**	**-2LL†**	**Nagelkerke**	**AIC††**	**BIC†††**
Step 1: Identify form of individual confounders						
PF (raw numbers, not banded, covariate)	1000	2	986.44	0.001	990.44	1000.26
PF (log, banded, treated as covariate)	1000	2	903.91	0.127	907.91	917.73
PF (banded, treated as factor)	1000	10	885.93	0.153	905.93	955.01
PF (log, not banded, treated as covariate)	1000	2	901.46	0.130	905.46	915.28
Decade (treated as covariate)	1000	2	931.79	0.085	935.79	945.61
Decade (treated as factor)	1000	7	915.78	0.109	929.78	964.13
Year (treated as covariate)	1000	2	930.02	0.088	934.02	943.84
Step 2: Identify form of joint confounders						
PF*Year	1000	5	904.55	0.126	911.728	921.544
PF, Year	1000	5	900.17	0.132	904.170	913.986

† 2 log likelihood

†† Akaike Information Criteria

††† Bayesian Information Criteria

5.2.1.2 Missing data, and new variable Morph-09

There are clear subgroups within the drug names, e.g. those that are not listed or not compliant with WHO guidelines, compared with those that are listed and compliant. The lack of information on STEM, for example, in those unlisted names means that measurements of the STEM do not apply to them. These groups are largely defined by these non-random patterns of missing data. This missingness may be mediated by the age of the name – older names are more likely to be not listed in the Stem Book, and to belong to smaller STEM taxa (that may later have been discontinued). These patterns of missing data will impact the analysis, by excluding cases with missing data, biasing the sample, by introducing apparent collinearity within groups, or by introducing bias by imputing the value of missing variables. Because of these risks to the analysis, the variable Morph-$08_{\text{taxon type}}$ was expanded to reflect the drivers behind missing data, such as not being listed in the WHO documentation (the Stem Book) or not being compliant with naming guidelines. The combined variable is Morph-$09_{\text{STEM taxon}}$. When analysed as part of Chapter 4, Morph-08 applied one of four possible values to a name, depending on the STEM it contained:

1. Single level - STEMS without a SUBSTEM
2. Multi-level - STEMS with a SUBSTEM
3. Complex - STEMS with two SUBSTEMS
4. Irregular - FREEFIX STEMS, STEMS that are not right-headed, SUBSTEMS without ‘parent’ STEMS

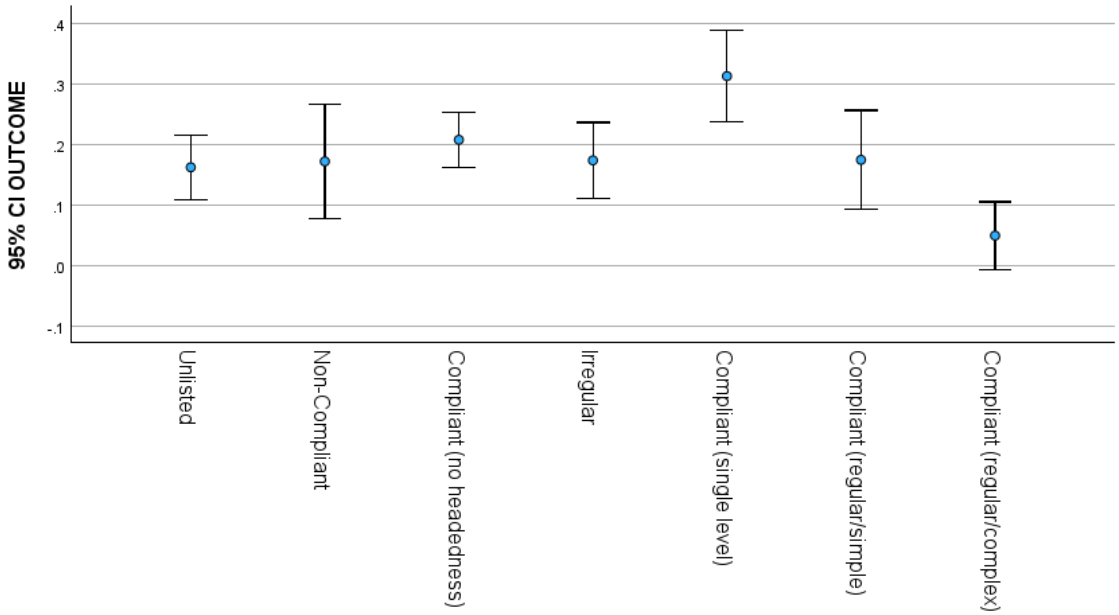
The first three values duplicate information from Morph-$02_{\text{contains a STEM}}$ and Morph-$03_{\text{contains a SUBSTEM}}$. The fourth value, ‘irregular’, duplicates information in Morph-$06_{\text{FREEFIX STEM}}$ and Morph-$07_{\text{right-headedness}}$. Morph-01 measures compliance with the naming guidelines in the Stem book, and includes a value for those names that are not listed in the Stem book at all. Older names that are taken to be well understood, such as paracetamol, are not listed, and this limits the information available on them for this analysis. These values are incorporated into the newly formed Morph-$09_{\text{STEM taxon}}$. Lastly, in Chapter 4 several variables were measures only for names in frequent STEM taxa – STEMS that appear at least five times in the dataset. This left around a third of the dataset of names that are listed in the Stem book and compliant, but nevertheless could not be analysed for semantic transparency or headedness because there simply was not enough information on them. The new variable Morph-09 has seven factors:

a. Unlisted [in the Stem book]
b. Not compliant [according to the Stem book]

c. Compliant, taxon has <5 members in dataset
d. Compliant, irregular type STEM [=>5 members]
e. Compliant, single-level type STEM [=>5 members]
f. Compliant, multi-level type STEM [=>5 members]
g. Compliant, complex type STEM [=>5 members]

Names that are not listed in the Stem Book, or listed as not being compliant with WHO guidelines, are associated with a lower risk of error than names that are listed and have information on their STEMS – see Figure 5.4.

Figure 5.4 Morph-09 STEM taxon, and error



5.2.1.3 Collinearity and structural overlap

There is considerable overlap in some of the variables because some are highly similar, some duplicates or subsets of others. For example, characters and syllables (Orth-01 and Orth-02) are closely related (Pearson rho=0.763, p<0.001). Because orthographic features are of interest here more than phonological features, characters are kept and syllables are removed. Rel-01 and Rel-03 are nested values of Rel-02 and Rel-04 respectively (Pearson rho=0.702, p 0.001; Pearson rho=0.804, p<0.001). For example, there are 312 names ending in the bigram -ne in the dataset. This is the sum of the frequencies of three-letter strings with ending in -ne: -ine (228 times), -one (64 times), -ane (12 times), and -ene (8

times). Since Rel-02 and Rel-04 provide more granular data, and because trigrams are used in other studies such as Shah et al. (2017), trigrams are kept (Rel-02 and Rel-04) and bigrams are removed (Rel-01 and Rel-03). Morph-04 is excluded because it is uninformative, with only a handful of positive values out of 1,000. Table 5.3 outlines which variables are kept or excluded from the analysis.

Table 5.3 Variables to keep or exclude

Variable	Exclude / keep	Reason for exclusion	Degrees of freedom (new)
	-		1
X01$_{\text{prescribing frequency}}$	Added		1
X02$_{\text{year}}$	Added		1
Orth-01$_{\text{characters}}$	Keep		1
Orth-02$_{\text{syllables}}$	Exclude	Removed due to collinearity with Orth-01	
Orth-03$_{\text{prohibited letters}}$	Keep		1
Orth-04$_{\text{Scrabble score}}$	Keep		1
Morph-01$_{\text{compliant}}$	Exclude	Duplication, so subsumed in new variable Morph-09.	
Morph-02$_{\text{includes STEM}}$	Exclude	Duplication, so subsumed in new variable Morph-09.	
Morph-03$_{\text{includes SUBSTEM}}$	Exclude	Duplication, so subsumed in new variable Morph-09.	
Morph-04$_{\text{includes variant STEM}}$	Exclude	Uninformative, almost no data.	
Morph-05$_{\text{incidental STEMS}}$	Keep		1
Morph-06$_{\text{includes FREEFIX}}$	Exclude	Duplication, so subsumed in new variable Morph-09.	
Morph-07$_{\text{right-headedness}}$	Exclude	Duplication, so subsumed in new variable Morph-09.	
Morph-08$_{\text{taxon type}}$	Exclude	Duplication, so subsumed in new variable Morph-09.	
Morph-09$_{\text{STEM taxon}}$	Added		6
Rel-01$_{\text{initial bigram frequency}}$	Exclude	Nested subset of Rel-02.	
Rel-02$_{\text{initial trigram frequency}}$	Keep		1
Rel-03$_{\text{terminal bigram frequency}}$	Exclude	Nested subset of Rel-04.	
Rel-04$_{\text{terminal trigram frequency}}$	Keep		1
Rel-05$_{\text{STEM frequency}}$	Keep		1
Rel-06$_{\text{ATC alignment broad}}$	Exclude	Nested subset of Rel-07.	
Rel-07$_{\text{ATC alignment}}$	Keep		1
Rel-08$_{\text{neighbourhood density}}$	Keep		1
Rel-09$_{\text{neighbourhood frequency}}$	Keep		1
Rel-10$_{\text{semantic transparency}}$	Keep		1
Rel-11$_{\text{top neighbour}}$	Keep		1

Of the remaining variables, a few have raised collinearity but within an acceptable level (Harlow, 2005; Licht, 1995; Sheskin, 2007). None have a Pearson correlation above 0.8, and the handful that are correlated, between 0.7 and 0.08, form an interrelated pattern (Dormann et al. 2013, Harlow, 2005; Licht, 1995; Sheskin, 2007). See Table 5.4. These are Morph-

$09_{\text{STEM taxon}}$, Rel-$05_{\text{STEM frequency}}$, Rel-$10_{\text{Semantic transparency}}$, and Rel-$07_{\text{ATC alignment}}$. This is because the three relational variables are calculated for a subset of the cases, only those names containing STEMS that have at least five members in the STEM taxon (45% of the dataset). The names with no values are missing and this aligns with the first three factors in Morph-$09_{\text{STEM taxon}}$. No further variables are removed at this stage.

Table 5.4 Collinearity between variables in the dataset: Pearson scores over 0.6.

	Morph-09	Rel-05	Rel-07	Rel-10
Morph-09	-			
Rel-05	0.701 **	-		
Rel-07		-0.705 **	-	
Rel-10	0.799 **	0.737 **		-

** significant at the 0.01 level

5.2.2 Variables for statistical analysis

After identifying confounders, excluding reductive or duplicating variables, and accounting for missing data, a global model was created including all candidate variables. Collinearity metrics were generated (variance inflation factor, VIF, and tolerance) for the full set and compared against the recognised threshold of VIF<10 (see Table 5.5). Some variables showed intermediate scores of between 5 and 10 but still fall within the acceptable range (Kim, 2019; Penn State, 2025). This was in addition to the Pearson's analysis described in the previous section, because this explores collinearity between variables selected for this statistical analysis, to be entered in the 'global model' and then subjected to variable selection. This is recommended in the SAMPL guidelines (Lang & Altman, 2013). The final dataset for analysis has 1,000 observations (INNs), one dependent variable (binary ERROR), and 15 independent variables, including prescribing frequency and year. Although these variables are described in the first instance in Chapter 4, descriptive statistics and frequencies are briefly given for these variables in Table 5.6 below because this dataset is a *subset* of the original dataset (excluding the 46 names without prescribing frequency data). Events per variable is 9.3 (=195/21), which is close to the recommended minimum of 10, so stability investigations (bootstrapping) of the results are performed (Heinze et al., 2018).

Table 5.5 Collinearity: tolerance and VIF scores

Variable	Tolerance	VIF
$X01_{\text{prescribing frequency}}$	0.911	1.098

X02$_{year}$	0.532	1.878
Orth-01$_{characters}$	0.769	1.300
Orth-03$_{prohibited\ letters}$	0.692	1.445
Orth-04$_{Scrabble\ score}$	0.856	1.168
Morph-05$_{incidental\ STEMS}$	0.615	1.625
Morph-09$_{STEM\ taxon,\ unlisted}$		
Morph-09$_{STEM\ taxon,\ non\text{-}compliant}$	0.563	1.776
Morph-09$_{STEM\ taxon,\ compliant\ small\ taxon}$	0.741	1.349
Morph-09$_{STEM\ taxon,\ irregular}$	0.427	2.341
Morph-09$_{STEM\ taxon,\ single\text{-}level}$	0.361	2.774
Morph-09$_{STEM\ taxon,\ multi\text{-}level}$	0.335	2.987
Morph-09$_{STEM\ taxon,\ complex}$	0.121	8.241
Rel-02$_{initial\ trigram\ frequency}$	0.795	1.257
Rel-04$_{terminal\ trigram\ frequency}$	0.801	1.248
Rel-05$_{STEM\ frequency}$	0.115	8.718
Rel-07$_{ATC\ alignment}$	0.375	2.664
Rel-08$_{neighbourhood\ density}$	0.792	1.263
Rel-09$_{neighbourhood\ frequency}$	0.802	1.247
Rel-10$_{semantic\ transparency}$	0.163	6.150
Rel-11$_{top\ neighbour}$	0.926	1.079

Table 5.6 Variables in the dataset for statistical analysis

Variable	Type of feature	Type of data	N valid cases [% of total cases]	Inclusion criteria	Descriptive statistics / frequencies			
					Minimum	***Maximum***	***Mean***	***Std deviation***
Prescribing frequency	Confounder	Continuous	1,000 [100.0]	All names	1.00	34.00	12.96	8.22
Year	Confounder				1955	2019	-	-
Orth-01$_{characters}$	Orthographic				5	23	10.61	1.83
Orth-04$_{Scrabble\ score}$	Orthographic				1	3.50	1.73	0.43
Morph-05$_{incidental\ STEMS}$	Morphosemantic				0	6	0.81	1.021
Rel-02$_{initial\ trigram\ frequency}$	Relational				1.00	18.00	3.74	3.96
Rel-04$_{terminal\ trigram\ frequency}$	Relational				1.00	228.00	71.09	85.01
Rel-08$_{neighbourhood\ density}$	Relational				0.00	25.00	4.35	4.33
Rel-05$_{STEM\ frequency}$	Relational		751 [75.1]	Listed and compliant in the Stem Book	0	63	10.07	16.26
Rel-07$_{ATC\ alignment}$	Relational		438 [43/8%]	Listed, compliant, in a STEM taxon with >5 names	-1.50	1.08	-0.57	0.46
Rel-10$_{semantic\ transparency}$	Relational				-0.81	1.81	0.81	0.64
					Yes [% valid cases]		***No [% valid cases]***	
LASA error	Outcome	Categorical, nominal [yes/no]	1,000 [100.0]	All names	195 [19.5]		805 [80.5]	
Orth-03$_{prohibited\ letters}$	Confounder				177 [17.7]		823 [82.3]	
Rel-09$_{neighbourhood\ frequency}$	Relational				441 [44.1]		559 [55.9]	
Rel-11$_{top\ neighbour}$	Relational				113 [11.3]		887 [88.7]	

					Unlisted	***Not compliant***	***Compliant, small taxon***	***irregular***	***Compliant, single-level***	***Compliant, multi-level***	***Compliant, complex***
Morph-09$_{STEM\ taxon}$	Morphosemantic, confounder	Categorical, factor [seven values]	1000 [100]	All names	185 [18.5]	64 [6.4]	313 [31.3]	144 [14.4]	147 [14.7]	86 [8.6]	61 [6.1]

5.3 Results

This section details the model fitting process, beginning with the assessment of all variables – including both linguistic predictors and non-linguistic confounders – within a comprehensive 'global model'. Following this, a stepwise method is applied to refine the model, resulting in a more parsimonious 'final model'. Finally, a bootstrapping procedure is presented, used to evaluate the stability and robustness of this final model. The results of these steps are reported and briefly discussed here, providing an overview of key patterns and trends. A full interpretation of these results is discussed in Section 5.4.

5.3.1 Fitting the model

The analysis dataset contains thirteen independent linguistic variables, two independent non-linguistic confounders, and one dependent variable. First, a 'global model' is built using all eligible predictors, linguistic and non-linguistic. All variables are included in the global model, with no selection for the global model, because it is intended as a reference. Results for the global model are given in Table 5.7.

Table 5.7 Global model

Global Model

	B	S.E.	Wald	df	Sig.	Exp(B)	95% CI
Constant	54.744	14.868	13.558	1	<0.001	5.96E+23	
X01$_{\text{prescribing frequency}}$	0.548	0.079	48.736	1	<0.001	1.731	[1.489, 2.026]
X02$_{\text{year}}$	-0.031	0.007	16.742	1	<0.001	0.970	[0.956, 0.984]
Morph-09$_{\text{STEM taxon, unlisted}}$			22.654	6	<0.001		[0, 0]
Morph-09$_{\text{STEM taxon, non-compliant}}$	-0.101	0.419	0.058	1	0.81	0.904	[0.385, 2.014]
Morph-09$_{\text{STEM taxon, compliant small taxon}}$	0.626	0.308	4.119	1	0.042	1.870	[1.028, 3.454]
Morph-09$_{\text{STEM taxon, irregular}}$	0.987	0.449	4.838	1	0.028	2.682	[1.107, 6.454]
Morph-09$_{\text{STEM taxon, single-level}}$	1.838	0.431	18.169	1	<0.001	6.282	[2.721, 14.811]
Morph-09$_{\text{STEM taxon, multi-level}}$	2.065	0.566	13.327	1	<0.001	7.884	[2.599, 24.013]
Morph-09$_{\text{STEM taxon, complex}}$	2.091	1.310	2.55	1	0.11	8.096	[0.633, 113.921]
Rel-02$_{\text{initial trigram frequency}}$	0.075	0.021	12.345	1	<0.001	1.078	[1.034, 1.124]
Rel-04$_{\text{terminal trigram frequency}}$	0.002	0.001	4.881	1	0.027	1.002	[1.000, 1.005]
Rel-10$_{\text{Semantic transparency}}$	-0.624	0.331	3.562	1	0.059	0.536	[0.278, 1.023]
Orth-03$_{\text{prohibited letters}}$	-0.511	0.280	3.348	1	0.067	0.600	[0.343, 1.029]
Rel-09$_{\text{neighbourhood frequency}}$	0.363	0.198	3.354	1	0.067	1.438	[0.975, 2.125]
Rel-07$_{\text{ATC alignment}}$	0.493	0.350	1.981	1	0.159	1.636	[0.813, 3.222]
Orth-04$_{\text{Scrabble}}$	0.238	0.219	1.18	1	0.277	1.268	[0.823, 1.942]
Rel-08$_{\text{neighbourhood density}}$	-0.016	0.023	0.467	1	0.494	0.984	[0.938, 1.029]
Rel-11$_{\text{top neighbour}}$	0.098	0.280	0.122	1	0.727	1.103	[0.628, 1.892]
Morph-05$_{\text{incidental STEMS}}$	-0.01	0.104	0.009	1	0.924	0.990	[0.806, 1.211]
Orth-01$_{\text{characters}}$	-0.01	0.053	0.033	1	0.857	0.990	[0.892, 1.099]
Rel-05$_{\text{STEM frequency}}$	-0.003	0.018	0.026	1	0.871	0.997	[0.960, 1.032]

After building the global model, backward elimination is used to obtain a 'final model'. For the final model, this process takes place only once; it is then repeated many times for the stability investigation. Backward elimination is used, with an exclusion rule based on Akaike information criteria (AIC, Heinze et al. 2018; Mantel, 1970). AIC is used as it corresponds to a p value of ≈ 0.157, where EPV is less than 25 (Heinze et al. 2018). As a consequence, some of the variables kept in the model are not significant to 5%, but they are all significant to 10% (however, individual categories within variables that are factors may not be below 10%, meaning that not all levels need to differ from each other for the factor variable to be a significant part of the model). It is important to note that the two confounders (prescribing frequency and year of approval), and the newly formed STEM taxon variable, Morph-09, were not subjected to variable selection. The confounders are both correlated strongly to the outcome and Morph-09 is considered to be integral to LASA error risk as it explains patterns of missingness in the other variables (Heinze et al. 2018). Therefore they were included in the model automatically and not subject to selection, to eliminate their effect on risk. All others may be interrelated and exchangeable when used for predicting LASA error, so were subject to elimination during the process. Steps taken in the backward elimination method are described in Table 5.8, with AIC scores. Each step removes the variable with the highest p value and stops when all p values are within the acceptable range of <0.157 and the AIC can be reduced no further. At each step, the AIC is reduced by approximately 2.

Table 5.8 Backward elimination model, steps

Step	Model	Degrees of freedom	-2 LL	Nagelkerke	Highest P value	AIC	BIC
Step 1	Global model with one dependent variable (error), two non-linguistic confounders (prescribing frequency and year) and thirteen linguistic predictors	20	815.982	0.250	0.871, Rel-05	855.982	954.137
Step 2	previous factors, less STEM frequency (Rel-05)	19	816.008	0.250	0.862, Orth-01	854.008	947.255
Step 3	previous factors, less Number of characters (Orth-01)	18	816.038	0.250	0.904, Morph-05	852.038	940.378
Step 4	previous factors, less Presence of incidental STEMS (Morph-05)	17	816.053	0.250	0.706, Rel-11	850.053	933.485
Step 5	previous factors, less Presence of a top-prescribed neighbour (Rel-11)	16	816.194	0.250	0.522, Rel-08	848.194	926.718
Step 6	previous factors, less Neighbourhood density (Rel-08)	15	816.612	0.249	0.273, Orth-04	846.612	920.228

IN FINAL MODEL - NO MORE REMOVED							
Step 7	previous factors, less Normalised Scrabble score (Orth-04)	14	817.803	0.248	0.096, Rel-07	845.803	914.512

The final model reduces the global model from thirteen to seven linguistic variables, and decreases the AIC in the global model from 856.0 to 845.8. The final model is given in Table 5.9 below. The adjusted pseudo-R^2 of 25% in the global model does not change in the final model (0.250 to 0.248). All variables in the final model are significantly correlated with the outcome *error* at the 10% level, and most at the 5% level. This includes the confounders, Morph-09$_{\text{STEM taxon}}$, Rel-02$_{\text{initial trigrams}}$, and Rel-04$_{\text{terminal trigrams}}$. The estimated shrinkage factor of the latter model is 0.87 and therefore additional post-estimation shrinkage is not necessary and would be easy to misinterpret due to correlations between variables (Dunkler et al., 2016; Heinze et al. 2018). This means that the stepwise modelling process inflated the parameter values (i.e. the coefficients should be around 87% as large).

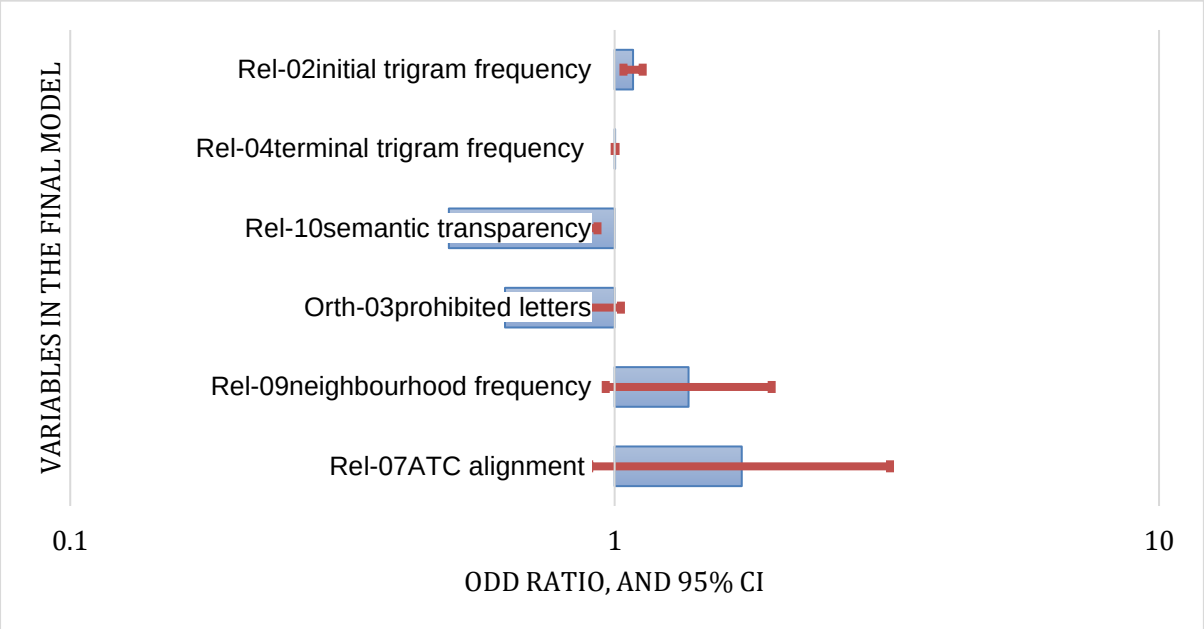
Table 5.9 Final model

	B	S.E.	Wald	Degrees of freedom	Significance	Exp(B)	95% CI
Constant	55.142	14.509	14.444	1	<0.001	8.86E+23	-
X01$_{\text{prescribing frequency}}$	0.548	0.078	48.967	1	<0.001	1.729	[1.489, 2.024]
X02$_{\text{year}}$	-0.031	0.007	17.671	1	<0.001	0.970	[0.956, 0.984]
Morph-09$_{\text{STEM taxon, unlisted (reference category)}}$	-	-	28.279	6	<0.001	-	-
Morph-09$_{\text{STEM taxon, non-compliant}}$	-0.128	0.415	0.096	1	0.757	0.879	[0.377, 1.944]
Morph-09$_{\text{STEM taxon, compliant small taxon}}$	0.599	0.268	5.005	1	0.025	1.820	[1.086, 3.110]
Morph-09$_{\text{STEM taxon, irregular}}$	0.948	0.381	6.184	1	0.013	2.581	[1.218, 5.452]
Morph-09$_{\text{STEM taxon, single-level}}$	1.892	0.396	22.884	1	<0.001	6.635	[3.078, 14.573]
Morph-09$_{\text{STEM taxon, multi-level}}$	2.110	0.529	15.888	1	<0.001	8.244	[2.917, 23.389]
Morph-09$_{\text{STEM taxon, complex}}$	2.204	0.929	5.628	1	0.018	9.064	[1.323, 53.316]
Rel-02$_{\text{initial trigram frequency}}$	0.078	0.021	14.237	1	<0.001	1.081	[1.038, 1.126]
Rel-04$_{\text{terminal trigram frequency}}$	0.002	0.001	5.741	1	0.017	1.002	[1.000, 1.004]
Rel-10$_{\text{semantic transparency}}$	-0.700	0.322	4.721	1	0.030	0.496	[0.261, 0.930]
Orth-03$_{\text{prohibited letters}}$	-0.463	0.253	3.35	1	0.067	0.629	[0.380, 1.027]
Rel-09$_{\text{neighbourhood frequency}}$	0.313	0.179	3.05	1	0.081	1.367	[0.963, 1.943]
Rel-07$_{\text{ATC alignment}}$	0.538	0.323	2.765	1	0.096	1.712	[0.898, 3.205]

Seven linguistic variables remain in the final model, including one that was 'forced' in so as to account for missing data. These are: Rel-02$_{\text{initial trigrams}}$; Rel-04$_{\text{terminal trigrams}}$; Rel-10$_{\text{semantic transparency}}$; Orth-03$_{\text{prohibited letters}}$; Rel-09$_{\text{neighbourhood frequency}}$; Rel-07$_{\text{ATC alignment}}$; Morph-09$_{\text{STEM taxon}}$. As

a whole, the unordered factor Morph-09$_{\text{STEM taxon}}$ is highly significant. These features of INNs appear to drive the risk of LASA error between names. Parameter estimates for these are shown in Figures 5.5 and 5.6.

Figure 5.5 Parameter estimates of variables in final model

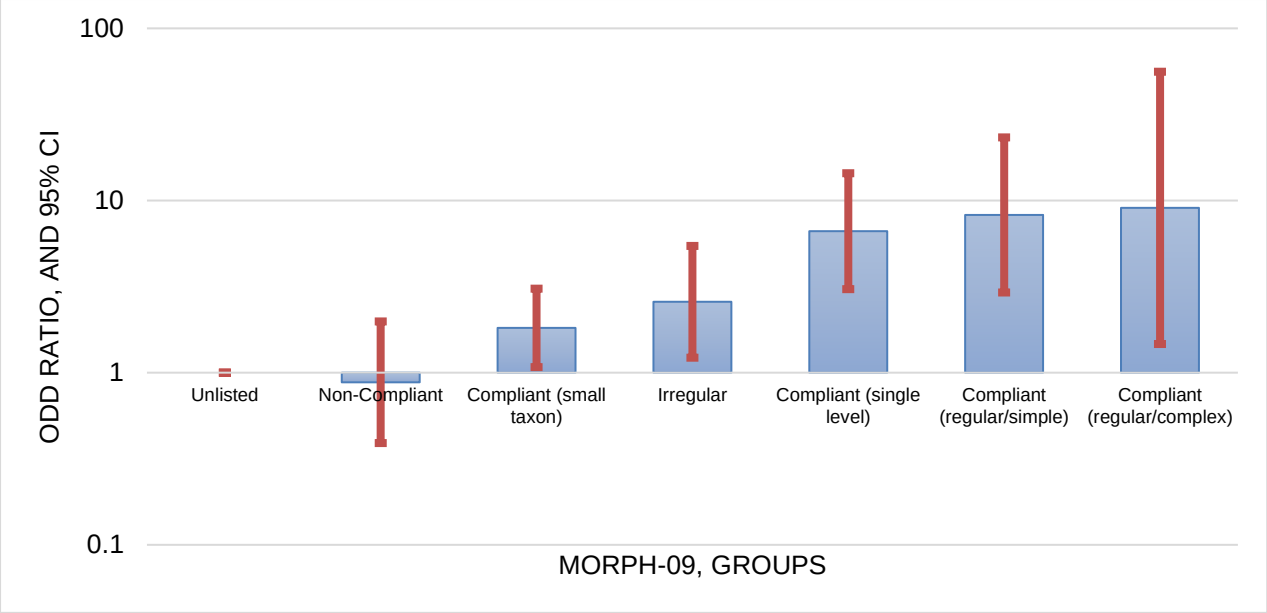


Having a frequent initial or terminal trigram (three-letter string at the beginning or end of the name) strongly increases the likelihood that the name will be involved in a LASA error (initial trigrams p<0.001, OR 1.081 [1.038, 1.126]) terminal trigrams p=0.017, OR 1.002 [1.000, 1.004]). Names with low semantic transparency, i.e. that do not show their pharmacological STEM in a predictable way, are also associated with higher LASA error incidence (p=0.030, OR 0.496 [0.261, 0.93]). Interestingly, names that *do not* contain prohibited letters have an increased risk of LASA errors (p=0.067 OR 0.629 [0.380, 1.027]). This was measured as a proxy for orthographic depth, so points to orthographic depth having a protective effect in this context. Having an orthographic neighbour that is prescribed more frequently has a facilitatory effect on LASA error (p=0.081 OR 1.367 [0.963, 1.943]), which agrees with findings in the literature (Lambert, Chang, & Gupta, 2003). Names that belong to a taxon that is well aligned with the ATC classification system are more likely to be involved in LASA errors (p=0.096 OR 1.712 [0.898, 3.205]), although the confidence interval is wide and this variable only appears in just over half of the bootstrapped resamples (50.4%), so the result should be interpreted with caution.

Two results are somewhat unexpected: increased ATC alignment (Rel-07) facilitating risk, and the presence of prohibited letters (Orth-03) having a protective effect. In this case it is prudent to look for the possibility of suppression of these variables by others in the model, however the odds ratios of these two variables do not change much between the global and final models (e.g. Rel-07, global model OR 1.636 p=0.159, final model OR 1.712 p=0.096), and indeed for any of the variables between the global and final models (see Tables 5.7 and 5.9).

Lastly, the new variable Morph-09$_{\text{STEM taxon}}$ was associated with significant variation in LASA error, with some differences between groups. The pattern of association is not as expected, with the last three groups showing increased association with LASA error (see Figure 5.5). These three groups are not significantly different to the other groups but not from each other. They are the compliant and regular STEM taxon types – they are compliant with WHO naming guidelines, right-headed, with no FREEFIX or variant STEMS. Irregularity in the STEM taxon of a name seems to have a protective effect against LASA errors.

Figure 5.6 Groups in Morph-09, STEM taxon



5.3.1.1 Accuracy of the final model

It is useful to explore the model's performance in distinguishing between high-risk names (names that have been reported in LASA errors) and low-risk names (names that have not been reported in LASA errors). The confusion matrix for the final model is given in Table

5.10. A prediction cutoff of 0.3 was used in the confusion matrix because in the case of LASA errors, a false negative can be costly: it means an error is missed. In the case of LASA error it is important to try to detect all possible high risk names, even if this leads to a small number of false positives.

The sensitivity, the model's ability to correctly identify high-risk names, is at an acceptable level of 52.8%. This suggests that the model can capture a significant portion of the high-risk names, but not all of them. This points to potential gaps in identifying crucial identifiers of high-risk names, thus increasing the risk that LASA-prone names will go undetected. At the same time, the model has a high specificity of 84.0%, indicating that it is very effective in correctly identifying low-risk names. However, because the dataset is imbalanced between negative and positive outcomes (80.5% no error, 19.5% error), most names are low risk, so it stands to reason that it was effective in capturing the error-free names.

The overall accuracy of the model stands at 77.9%, indicating its overall efficacy in predictions. The positive predictive value (PPV), at 44.4%, means that when the model predicts a name as high-risk, it is correct almost half the time. This implies room for improvement in terms of reducing false positives - making predictions in complex data is, however, challenging and performance tends to be poor. The F1 score, considering both precision and sensitivity, stands at 48.2%, indicating a relatively low balance between the model's ability to identify both high-risk and low-risk names effectively and significant room for improvement in predicting high-risk names.

In summary, while the model is highly effective in accurately identifying low-risk names, it struggles to capture high-risk names. The default prediction is low risk because the average risk is low (just under 20% of the dataset). Most likely there are other important clinical factors that have an impact, but they are not included in this dataset or factored into the model, which might explain why the sensitivity in detecting high-risk names is not higher.

Table 5.10 Confusion matrix for final model

		Predicted LASA error*	
		No	Yes
Reported LASA error	No	676	129
	Yes	92	103

* cutoff 0.3

5.3.2 Bootstrapping

Assessing robustness of a final model is an important part of the analysis process and is recommended in the guidelines followed here (Heinze et al. 2018). The model building process that was used to construct the final model was repeated using the resampled data; for each of the 1,000 bootstrapped samples the final model was recorded, including model coefficients. These models were used to assess the robustness of the final model.

Model selection frequency is a useful way to assess robustness of the final model compared to the resamples, measuring how often the final model appears compared to other possible models. In this case it performs well - the final model is stable as it was selected in 3.1% of the resamples (31/1,000) and critically, it is the *most selected model* in the resamples. With a large number of variables precise individual models only appear quite rarely, so 3.1% indicates the model performs well. The top 20 models are given in Table 5.11 and 'Bootstrap inclusion frequency' is given in Table 5.12, with other stability measures that are discussed later in this section. The top few models are virtually identical, and most of the linguistic variables in the final model are in the top twenty models too (e.g. Rel-02$_{\text{initial trigrams}}$, Rel-04$_{\text{terminal trigrams}}$, Orth-03$_{\text{prohibited letters}}$ are all in the top five models).

Table 5.11 Top twenty models in the bootstrap resamples

Included predictors (variable names)	Frequency	% of ALL	cum %
O03 R02 R04 R07 R09 R10	31	3.1%	3.1%
O03 R02 R04 R09 R10	28	2.8%	5.9%
O03 O04 R02 R04 R09 R10	20	2.0%	7.9%
O03 R02 R04 R07 R10	19	1.9%	9.8%
O03 O04 R02 R04 R07 R09 R10	17	1.7%	11.5%
O03 R02 R04 R10	17	1.7%	13.2%
O03 R02 R04 R07 R08 R09 R10	15	1.5%	14.7%
O03 O04 R02 R04 R07 R10	13	1.3%	16.0%
O03 R02 R04 R07 R09	13	1.3%	17.3%
O03 O04 R02 R04 R05 R09	12	1.2%	18.5%
O03 O04 R02 R04 R10	10	1.0%	19.5%
R02 R04 R07 R09 R10	9	0.9%	20.4%
R02 R04 R09 R10	9	0.9%	21.3%
O03 O04 R02 R04 R07 R08 R09 R10	8	0.8%	22.1%
O03 O04 R02 R04 R07 R09	8	0.8%	22.9%
O03 R02 R04 R07 R09 R10 R11	8	0.8%	23.7%
O03 R02 R04 R09	8	0.8%	24.5%
M05 O03 O04 R02 R04 R09 R10	7	0.7%	25.2%
O03 R02 R04 R07 R08 R09	7	0.7%	25.9%

O03 R02 R04 R08 R09 R10 R11	7	0.7%	26.6%
**** final model***			

Legend:
O03: prohibited letters
R02: initial trigrams
R04: terminal trigrams
R05: STEM frequency
R07: ATC alignment
R09: neighbourhood frequency
R10: semantic transparency
O04: Scrabble
R08: neighbourhood density
R11: top-prescribed neighbour
M05: incidental STEMS

Median estimates are a useful way to check that the effect sizes of the variables in the final model are comparable to their effect sizes in the resamples. As the mid-point in the effect size of the variable across all models, if the observed coefficient in the final model is close to the median this shows that the effect observed is in line with all similar models. Table 5.12 shows they are comparable. Median estimates are the midpoint in the range of fitted values for the coefficient in the final model. This includes cases where the variable is excluded, in which case it is coded as a zero. Where the median estimate is zero this is likely due to lower inclusion frequency. The model suggested by the bootstrap medians is in line with the final model (those that are present in at least half of the resamples).

Heinze et al. (2018) recommend further assessing the robustness of the final model by calculating the root mean squared difference (RMSD) ratio. Essentially the RMSD ratio assess how robust the standard error of the estimate is by measuring the potential variance of the estimate in practice. In this case, the RMSD of almost all variables is close to 1 which means the measured SE is likely fairly robust. But those linguistic variables with bootstrap inclusion frequency of less than ~80% are further away from zero, indicating less precision in the estimates. This indicates that the confidence intervals stated in the final model are narrower than perhaps they should be, because they are based on only one model instead of many.

Similarly to the RMSD, relative conditional bias measures the effect the selection process has on the estimate itself. In other words, is the estimate in the final model likely to be biased up or down? For most of the variables in the final model the RCB is small, less than 10%, with only one exceeding 50%. The bias is negligible for most of the factors of Morph-$09_{\text{STEM taxon}}$ and for Rel-$02_{\text{initial trigrams}}$, which has a bootstrap inclusion frequency of 97.9%, but it increases to 30% in the other selected linguistic variables, where variable selection is less

certain. The RMSD ratio and relative conditional bias are given in Table 5.12. The coefficients of the selected variables are very similar to the global model estimates, indicating no selection bias in the final model. The 5% and 95% percentiles can be regarded as confidence intervals obtained through resampling-based multi-modal inference.

Table 5.12 Bootstrapping of final model

Global model				Selected model						
Predictors	**Estimates**	**Standard error**	**Bootstrap inclusion frequency (%)**	**Estimate**	**Standard error**	**RMSD ratio**	**Relative cond. bias (%)**	**Bootstrap median**	**5th**	**95th**
(Intercept)	54.744	14.868	100	55.142	14.509	1.00	+0.9	55.532	30.733	79.437
X01 prescribing frequency	0.548	0.079	100	0.548	0.078	1.03	+2.8	0.561	0.439	0.702
X02 year	-0.031	0.007	100	-0.031	0.007	1.07	-0.4	-0.031	-0.043	-0.019
M09 STEM taxon, unlisted			100							
Morph-09 STEM taxon, non-compliant	-0.101	0.419	100	-0.128	0.415	1.04	+16.5	-0.102	-0.863	0.584
Morph-09 STEM taxon, compliant small taxon	0.626	0.308	100	0.599	0.268	1.05	+4.3	0.669	0.127	1.172
Morph-09 STEM taxon, irregular	0.987	0.449	100	0.948	0.381	1.07	-7.7	0.913	0.145	1.707
Morph-09 STEM taxon, single-level	1.838	0.431	100	1.892	0.396	1.09	+0.6	1.829	1.119	2.669
Morph-09 STEM taxon, multi-level	2.065	0.566	100	2.11	0.529	1.04	-1.5	2.067	1.055	2.981
Morph-09 STEM taxon, complex	2.091	1.310	100	2.204	0.929	2.44	-31.5	1.892	-0.839	4.263
Rel-02 initial trigram frequency	**0.075**	**0.021**	**97.9**	**0.078**	**0.021**	**1.13**	**+4.4**	**0.077**	**0.041**	**0.115**
Rel-04 terminal trigram frequency	**0.002**	**0.001**	**80.5**	**0.002**	**0.001**	**1.36**	**+40.1**	**0.002**	**0.000**	**0.004**
Rel-10 semantic transparency	**-0.624**	**0.331**	**73.4**	**-0.700**	**0.322**	**1.34**	**+37.1**	**-0.685**	**-1.327**	**0.000**
Orth-03 prohibited letters	**-0.511**	**0.280**	**70.7**	**-0.463**	**0.253**	**1.24**	**+29.0**	**-0.535**	**-0.991**	**0.000**
Rel-09 neighbourhood frequency	**0.363**	**0.198**	**64.0**	**0.313**	**0.179**	**1.33**	**+29.2**	**0.341**	**0.000**	**0.699**
Rel-07 ATC alignment	**0.493**	**0.350**	**50.4**	**0.538**	**0.323**	**1.21**	**+54.2**	**0.434**	**0.000**	**1.077**
Orth-04 Scrabble	0.238	0.219	39.0					0.000	0.000	0.622
Rel-08 neighbourhood density	-0.016	0.023	26.8					0.000	-0.058	0.000
R11 top neighbour	0.098	0.28	23.1					0.000	-0.452	0.616
Morph-05 incidental STEMS	-0.01	0.104	20.6					0.000	-0.221	0.178
Orth-01 characters	-0.01	0.053	19.7					0.000	-0.105	0.081
Rel-05 STEM frequency	-0.003	0.018	17.3					0.000	-0.039	0.000

5.3.2.1 Relationships between variables in the final model

Pairwise inclusion frequencies inform about connections between independent variables, so-called 'rope teams' and 'competitors', by looking at how often pairs of variables in the final model appear and whether it is more or less often than would be expected based on their selection in the bootstrap resamples. It sheds light on collinearity and suppression but also highlights pairs of variables working together. This is shown in Table 5.13. For example, characters and prohibited letters were selected in 10.7% of the resamples, while one would expect a frequency of 13.9% (=19.7% x 70.7%). Therefore the pair is marked with a "-" in the lower triangle of Table 5.13 – it is *underrepresented*. This means that the departure from the expected value is significant. This is explained by them performing the same role so excluding each other from models when they are competing. They vie for inclusion and can represent similar information or highly correlated predictors. On the other hand, neighbourhood density and neighbourhood frequency are marked with a "+" in Table 5.12 because they are simultaneously selected in 22.0% of the resamples, while the expectation under independency is 17.2% (=26.8% x 64.0%). They co-occur more than expected – they are *overrepresented*. This indicates that they share some information and are useful together in a model. In the table, significance of $\chi 2$ at the 0.01 level was used as the formal criterion for the flags. Semantic transparency, ATC alignment, and STEM frequency also form 'rope team'. Most of interest, prohibited letters form a rope team with Scrabble score and terminal trigrams, but characters is a competitor to prohibited letters. It may be that the prohibited letters measure provides complementary information on orthographic complexity both in endings and throughout names, but provides sufficient information on name length to push characters out of the model. Unexpected pairwise inclusion frequency points to collinearity, and the VIF scores for all variables in the global model show come collinearity, but at an acceptable level (Kim, 2019; Penn State, 2025, see Table 5.5).

Table 5.13 Pairwise inclusion frequency

	Morph-05	Orth-01	Orth-03	Orth-04	Rel-02	Rel-04	Rel-05	Rel-07	Rel-08	Rel-09	Rel-10	Rel-11
Morph-05		44	145	86	201	155	33	97	62	135	151	44
Orth-01			107	73	194	155	38	108	68	122	131	44
Orth-03		-		301	688	587	123	348	188	468	503	168
Orth-04			+		381	312	73	188	105	249	275	96
Rel-02						789	169	493	264	632	722	226
Rel-04			+				147	415	202	522	586	182
Rel-05								48	42	107	110	29
Rel-07							-		141	326	347	118
Rel-08		+								220	188	74
Rel-09									+		472	137
Rel-10							+	+				162
Rel-11												

5.4 Discussion and conclusions

Seven linguistic features are identified as significantly associated with LASA error in the final model: Rel-02$_{\text{initial trigrams}}$; Rel-04$_{\text{terminal trigrams}}$; Rel-10$_{\text{semantic transparency}}$; Orth-03$_{\text{prohibited letters}}$; Rel-09$_{\text{neighbourhood frequency}}$; Rel-07$_{\text{ATC alignment}}$; Morph-09$_{\text{STEM taxon}}$. Both non-linguistic confounders, prescribing frequency and year of approval, are both strong predictors of risk and must be taken into account when assessing the risk of LASA errors and included in any analytical framework. Prescribing frequency significantly affects models, with a notable impact size ($p<0.001$, Nagelkerke pseudo R-squared of 12.7% alone), as does year of approval ($p<0.001$, Nagelkerke 8.5%).

Names with a high number of counterparts with the same beginning or end strings have a high level of shared features, which makes them more likely to be confused. This partly confirms findings in the literature, that high-frequency *beginning strings* correlate with error (Shah et al., 2017), but additionally demonstrates that high-frequency *end strings* are also correlated with error. Despite weighting the method for calculating orthographic neighbourhood to the beginning of names, neighbourhood density was not found to be a significant correlate of LASA errors, suggesting that orthographic similarity at the beginnings and endings of names is a more accurate predictor of error than the orthographic similarity between full names.

Semantic transparency was also associated with error – names that display their STEM in a predictable position (i.e. always at the end) have a lower risk of LASA error. For example, the STEM taxon -conazole, with ten members in the dataset, e.g. econazole, has a high overall semantic transparency (all over 1.04 PFS ratio) but a low error rate (1/10 members involved in a LASA error: fluconazole). As an example of the reverse, -azepam has low overall semantic transparency (all under 0.2 PFS ratio) but a high error rate (8/13 members involved in a LASA error, e.g. diazepam). Variant STEMS were included here with STEMS that were in a different position, so this measure incorporates other aspects of unpredictability. This finding suggests that semantic transparency is associated with a change in LASA error risk, complementing the body of research on semantic transparency in English general language compounds but this result should be interpreted with caution, since the stability investigations suggest that its effect is biased upwards (Bell & Schäfer, 2016; Libben, 1998; Libben et al., 2003).

Prohibited letters were significantly associated with LASA error, but contrary to expectation the presence of such letters has a protective effect. The variable is used as a proxy for

orthographic depth, i.e. the correspondence between grapheme and phonemes in a linguistic unit. Many of the prohibited letters exacerbate orthographic depth, e.g. <ph> for /f/ and <y> for /i/, which will in turn reduce accurate perception of names. It could be that this is not an appropriate proxy measure, since in many cases the prohibited letters do not necessarily inhibit orthographic depth, because they represent a phoneme, e.g. <th> for /ð/ or /þ/. It may also be that the use of digraphs such as <ph> add to distinctiveness of the names and trigger recognition and recall. For prohibited letters, it is likely that their effect in the final model is confounded by the effect of year. Older names are strongly associated with increased risk of error ($p<0.001$, OR 0.970 [0.956, 0.984]) and prohibited letters are most prominent in older names, approved in the early days of the INN programme (63% of names containing a prohibited letter were approved before 1960). Their presence drops off steeply from the seventies, showing increased compliance with this WHO guideline. Alone, names containing prohibited letters are involved in more LASA errors than those not containing them (30% with, compared with 23% without) so a protective effect against error is not demonstrable.

Neighbourhood frequency is significantly associated with LASA error: names with a more frequently prescribed neighbour are more likely to be involved in a LASA error. This is despite not having an effect in studies conducted by Lambert et al. (Lambert et al., 2001b; Lambert, Chang, & Gupta, 2003). This is likely due to a difference in definition: Lambert measured the mean prescribing frequency of the neighbourhood compared with the target name, whereas here the variable measured whether *any* of the names had a higher prescribing frequency than the target name. However, this result should be treated with caution – the confidence interval is broad and it spans 1, meaning that the direction of the effect is unclear. It is not significant to 5% in the final model ($p=0.081$) and the stability investigations suggest that its effect is biased upwards.

ATC alignment is also significantly associated with LASA error, yet unexpectedly the results point to increased alignment having a protective effect, though the confidence intervals are wide so this finding is uncertain. With a closer look, there is likely some shared function between $\text{Rel-10}_{\text{Semantic transparency}}$ and $\text{Rel-07}_{\text{ATC alignment}}$. Both variables measure uniformity within a STEM taxon: semantic transparency is uniformity in the position of the STEM; ATC alignment is uniformity of ATC code. They are correlated (Pearson rho=-0.603, $p<0.001$) and form a so-called *rope team* in the bootstrap resamples in appearing together more than expected. It suggests that internal harmony and consistency of STEM taxa creates predictability for users, which in turn reduces the risk of LASA error. The effect is uncertain for ATC alignment: despite the positive effect size, the confidence interval is large (OR 1.712

[0.898, 3.205]) and it spans 1, so the direction of the effect is unclear. It is not significant in the global model and is not significant to <5% in the final model (p=0.096).

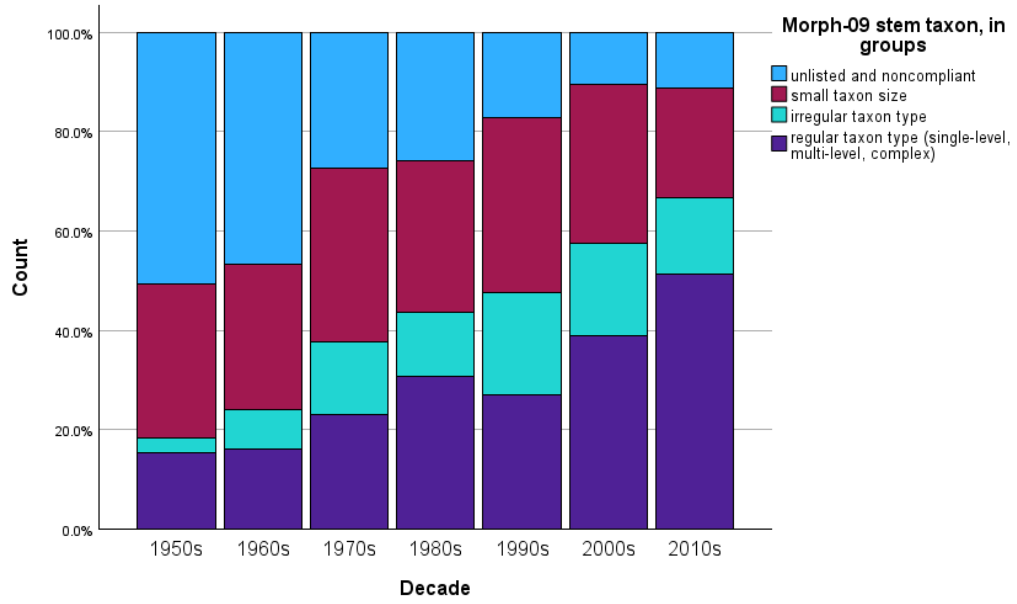
Finally, the newly formed variable Morph-09$_{\text{STEM taxon}}$ was significantly associated with LASA error. As described earlier in the chapter, the last three groups form a cluster that has an increased incidence of LASA error. These are the compliant and regular STEM taxon types: they are all right-headed, contain no FREEFIX or variant STEMS, and have well-defined STEM and SUBSTEMS. It may be that some aspects of irregularity (e.g. a FREEFIX STEM, left-headedness, or variant STEM), prompt the user to pay more attention to the name and this has a protective effect against error. In comparison, unlisted and non-compliant names are associated with low LASA error incidence. The names in these groups are among the oldest in the dataset and those with the highest prescribing frequency (e.g. paracetamol). Since Morph-09$_{\text{STEM taxon}}$ was not subjected to variable selection, confounding effects of prescribing frequency and year may be overinflating the effect of STEM taxon type.

The ‘complex’ category is wholly comprised of monoclonal antibodies with the dual SUBSTEM system; however, in 2017 one of the SUBSTEMS was removed from newly designated monoclonal antibodies, and since 2022, the -mab STEM has been replaced by four separate STEMS: -tug; -bart, -mig, and -ment (Balocco et al., 2022). So, INNs representing monoclonal antibodies are complex type 3 if they were approved before 2017 (e.g. trastuzumab with SUBSTEMS -tu- and -zu-), multi-level type 2 if approved between 2017 and 2022 (e.g. erenumab with SUBSTEM -u-), and single-level type 1 if approved after 2022 (e.g. epacmarstobart, with STEM -bart). This illustrates the dynamic nature of pharmaceutical nomenclature - guidelines change over the years but this does not affect names retroactively and names created under now-defunct guidelines persist in the nomenclature and contrast with their younger counterparts. The ‘complex’ group of Morph-09 appears to be associated with increased risk of error, but the finding is unreliable – although it is significant, the effect is heavily biased (relative conditional bias of -31.5%) and the confidence intervals are very wide (1.323, 53.316).

The taxonomy of names has become more complex over the years. The use of single-level taxa names (names with a single STEM to denote the pharmacological group, with no SUBSTEMS) has declined since the 1980s, with multi-level taxa (names with a STEM and a SUBSTEM) on the rise. From the 1970s onwards, there is a material shift in naming, most probably because of revision in WHO naming policy and maturation from the growing pains of the programme’s early years. Names approved in the first two decades of the programme are longer and contain multiple letters that are now prohibited by guidelines (e.g.

phenoxymethylpenicillin). They are also some of the most prescribed medications, e.g. paracetamol, with 1.2 billion items prescribed annually in England and Wales. Fewer of these older names are listed in the Stem Book so could not be analysed in this study for their use of STEM and their taxon type. Names prescribed in the 1950s and 1960s make up 13.7% of the dataset (137/1000 names) and 33.3% of names with a reported LASA error (65/195). Figure 5.7 shows the change over time of these characteristics, grouping together 'unlisted and non-compliant', and the 'regular' types of STEM taxon. Regular types are aggregated because the complex taxon type of names for monoclonal antibodies no longer exists, and no other names have been assigned a double SUBSTEM, so in future, regular STEM taxa will be either single- or multi-level. As the figure shows, the 'regular' taxa are becoming more commonplace, and taxa are getting larger, with fewer names in a small taxon in recent decades.

Figure 5.7 Morph-09 STEM taxon, in groups



5.4.1 Limitations

This statistical analysis has several key limitations that affect the depth and accuracy of the findings. Firstly, linguistic variables were explored for the English-language nomenclature only (described in Chapter 4), and not in other languages, because information on pharmacological STEMS is available for English-language INNs only. Secondly, LASA error reports were available in only English so statistical analysis (described in this Chapter 5) was not possible for INNs in languages other than English. Thirdly, the dataset of INNs was restricted to names of medications that are authorised for use in the UK, which excluded

over 80% of the full INN nomenclature (see Section 4.2.1 for primary data collection) – this was because names that are not authorised have a zero possibility of being involved in an error, and are not interesting from a clinical perspective.

The INN nomenclature in this dataset has been formed over more than seven decades (between 1955 and 2019) and incorporates diachronic changes in naming and prescribing patterns, which complicates analysis and potentially introduces biases. This analysis was furthermore carried out on a subset of the nomenclature: names for medications that are authorised in the UK, and one-word names. Multi-word names were excluded, as in the previous study in Bryan et al. (2015), to allow for comparisons based on factors like name length. They cannot be fairly compared with single-word names and may skew results. Furthermore, many multi-word names contain modifiers (e.g. “sodium” or “acid”), making them longer and more syllabic but not necessarily more complex, and these modifiers are often dropped in usage.

The variables analysed in this chapter are linguistic features of names, and do not encompass all the possible driving factors of LASA error. Non-linguistic factors such as stress levels, urgency of situations, and contextual factors remain unaccounted for, potentially influencing error rates. The outcome measure is binary, because detailed data on LASA errors was not available. This precludes a nuanced understanding of the drivers of LASA errors, because there is no distinction between names that have been confused a thousand times and names that have been confused only a handful of times. More granular data are needed, such as the frequency and seriousness of errors, and a detailed exploration of variables like the clinical setting of the error, consequences of the error, and first language of the healthcare professional. Furthermore, the measurement of a drug name’s neighbourhood was not restricted to a clinical specialty so as to capture all names in a neighbourhood.

The method used for the statistical analysis was stepwise variable selection using backward elimination. Firstly, using stepwise methods risks overfitting the training data, leading to poor performance in the validation stage, and a false sense of objectivity, in that the model is optimised but may not include all necessary variables and may not translate to real-life scenarios. Secondly, the presence of non-random missing data affected the analysis by the possibility of introducing bias, and a high number of correlated variables makes it hard to distinguish between direct and indirect effects, making any interpretation of findings more challenging. Thirdly, some low levels of collinearity remained in the global model. Fourthly, the ‘events per variable’ (EPV) was high, which was counteracted by a series of

bootstrapping investigations. Finally, two non-linguistic confounders (prescribing frequency and year of approval) dominate the risk of LASA errors in the statistical analysis. This means that even minor imperfections in the way in which these confounders have been modelled may have large consequences for the final combined models.

Chapter 6: Pharmaceutical nomenclatures in Russian and Italian

6.1 Introduction

This chapter explores pharmaceutical nomenclatures in Russian and Italian and discusses the implications of the findings in Chapter 5 for multilingual pharmaceutical nomenclature, using two case study languages of Russian and Italian. This answers Research Question 3 (RQ3): *How can Russian-language and Italian-language pharmaceutical nomenclatures be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?* In this chapter, linguistic features of the pharmaceutical nomenclature in Russian and Italian will be explored and described. This follows the linguistic analysis of the English-language nomenclature in Chapter 4. It includes a consideration of how features associated with LASA errors (identified in Chapter 5) change in translation and in multilingual nomenclatures. The case studies contain references to linguistic concepts and theory that are explained in Chapter 2.

Methods for this chapter are presented in Section 6.2, including a breakdown of which linguistic features can be analysed in other languages. A full linguistic analysis was not available in any languages other than English, because naming guidelines are published by the WHO solely for the English nomenclature and not for the other languages in which INNs are assigned. The WHO INN Stem Book, which lists all pharmacological STEMS and SUBSTEMS for INNs, is published solely in English, for English-language INNs. Likewise, statistical analysis between linguistic features of INNs and LASA errors was only possible in English because any LASA error reports identified were in the English language (though it is not clear if the reported errors occurred in an English-language setting). Therefore, owing to a lack of guidance on the STEM system, and a lack of data on error reports, Russian and Italian are included in RQ3 only. This suits the sequential nature of the research questions: RQ2 uses the findings of RQ1, and RQ3 uses the findings of RQ2.

The scope of work presented is limited by the author's own language proficiency – a native speaker of English, fluent in Italian, with basic proficiency in Russian, Spanish, and Welsh. Italian and Russian are selected as case-study languages as they represent differential points of interest: Italian is a Romance language, using the Roman alphabet, with a national naming system outside of the WHO nomenclature; Russian is a Slavic language, using the Cyrillic alphabet, and Russian INNs are translated by the WHO (because Russian is one of the official working languages of the United Nations).

This chapter considers the implications of findings from Chapter 5 for pharmacological nomenclature in other languages. By adopting a case study approach it will explore and describe the linguistic features of names in Russian and Italian. Each case study will explore the nomenclature in that language and compare with English, where possible. These two case studies are designed to be read in isolation, so there is inevitably some overlap between them, e.g. some figures appear in both.

As in previous chapters, INNs are in blue and official STEMS are in green. When a STEM is the focus of discussion, it may be highlighted in the name, e.g. tamoxifen. Russian-language INNs are in purple and Italian Common Names are in orange. Phonemes are given in /forward slashes/, graphemes are given in <angle brackets>, foreign-language words are in *italics*, and glosses of foreign words are given in 'single quotes'. Nonce words, including hypothetical medication names, are given an *asterisk.

6.2 Data analysis methods for RQ3: Multilingual implications

To answer RQ3, linguistic features explored as part of Chapter 4 are described for the Italian and Russian nomenclatures, where the availability of data allows, organised into orthographic, morphosemantic, and relational features. A case-study approach is adopted to explore nomenclatures in Russian and Italian, and features of the names are presented using descriptive statistics and complemented by a list of STEMS.

6.2.1 Linguistic features described for the Russian and Italian nomenclatures

It was not possible to analyse all of the linguistic features in Russian and Italian that were analysed as part of Chapter 4 in English. This is because the WHO official Stem Book is published in English only, for English-language names only. Error reports were not identified in Russian and Italian in any literature, and so this variable was also restricted to English (see search strategies outlined in Chapter 3). Because the WHO doesn't publish naming or

STEM guidance in Russian or Italian, it was not possible to explore: the presence of prohibited letters such as <k> or <y> (measure Orth-03); positional family size ratio (the predictability of the semantic category based on the position of the STEM, measure Rel-10); the frequency of the STEM (Rel-05); or indeed any of the eight morphosemantic features (Morph-01 – Morph-08).

For some variables, it would have been possible to collect data for other languages, but that presented a large time burden for little return. For example, prescribing frequency data are available on request for Italian medicinal products approved by the European Medicines Agency, but since no LASA error data are available from Italy the prescribing frequency would not serve a purpose (as a confounder for statistical analysis) and prescribing frequency in the UK is likely to be largely extrapolable for Italy. Without prescribing data, neighbourhood frequency cannot be measured (Rel-09). Likewise, the year in which the name was approved may be slightly later for Italian names than for English and Russian INNs (since the INN is published first, with national naming agencies following) but by using decade this variability will be reduced. The number of syllables was not measured, because this correlates closely with the number of characters (Pearson rho=0.773, p<0.001). Other variables do not change in other languages, such as ATC codes (measures Rel-06 and Rel-07). ATC codes are alphanumeric codes assigned by the WHO to medicinal products, and are used internationally.

This leaves seven variables that can be explored for both Russian and Italian nomenclatures: the number of characters in the name (Orth-01); the Scrabble score, using country-specific scores (Orth-04); all ngram frequencies, which measure the frequencies of the strings at the beginnings and endings of names across the dataset (Rel01-04); and neighbourhood density, which measures the number of names to which a name is orthographically similar (Rel-08). In line with the three types of linguistic features presented in Chapter 4, for Russian and Italian names two orthographic features, no morphosemantic features, and five relational features are analysed. The linguistic features and confounders that are analysed for Russian and Italian are listed in Table 6.1.

Table 6.1 Linguistic features and confounders that can be explored in Russian and Italian

ID	Variable	Type of feature	Type of data
Orth-01	How many characters does the INN contain?	Orthographic	Continuous
Orth-04	What would the INN score be if played in a Scrabble game?	Orthographic	Continuous

Rel-01	How frequently do the first two letters of the name (initial bigram) appear in the data set?	Relational	Continuous
Rel-02	How frequently do the first two letters of the name (initial trigram) appear in the data set?	Relational	Continuous
Rel-03	How frequently do the first two letters of the name (terminal bigram) appear in the data set?	Relational	Continuous
Rel-04	How frequently do the first two letters of the name (terminal trigram) appear in the data set?	Relational	Continuous
Rel-08	Neighbourhood density: how many names are similar to the INN?	Relational	Continuous

These variables were calculated in the same way as their English counterparts, as described in Section 5.3. The results are presented in Chapter 8. None of the morphosemantic features can be formally explored, because no Stem Book is published for languages other than English, so it is not possible to identify official STEMS (i.e. STEMS that are defined and used by the WHO INN programme) in Russian or Italian. However, the case studies in Chapter 8 will attempt to describe how STEMS are translated and transliterated in Italian and Russian, and to present a list of common STEMS in these languages.

Russian uses the Cyrillic alphabet, so transliteration mechanisms are also explored as part of the Russian case study in Chapter 8. ISO 9:1995 is the official transliteration protocol for Cyrillic > Latin transliteration used internationally, so to test the compliance to the protocol the Russian name was transliterated into English using the protocol, then measured the difference between the back-transliterated name and the original English name, to identify where the transliteration errs from the standard. The WHO does not stipulate whether they use this transliteration standard or another one, but this is the international standard and therefore a likely candidate. The results are discussed in Chapter 8.

6.3 Case study 1: Russian-language INNs

6.3.1 Russian-language INNs: overview

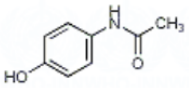
This case study explores linguistic features of Russian-language INNs (***международное непатентованное наименование***, Pron. ‘mezhdunarodnoye nepatentovannoye naimenovanye’) and the translation and transliteration of English-language INNs into Russian. The Russian abbreviation would be *МНН* but for ease of reading the terms *Russian-language INNs* and *INNs* are used to distinguish between Russian and English names. Where possible, the case study compares features between Russian and English. INNs are published first in English, and then translated into the other five official languages

of the United Nations (UN; Arabic, Chinese, French, Russian, and Spanish), and additionally Latin (SoINN, 2020b).

English is used as the *de facto* anchor language and source for translation into the other UN languages (SoINN, 2020b), and more broadly into other languages in other national nomenclatures. In translation discourse, English is the 'source' language and Russian is the 'target' language. These languages have been used since the inception of the INN programme in 1951, so for each English name, there is a translation of the name into each of the UN languages. For example, paracetamol is published on Mednet, the online INN application portal, as in Figure 6.1.

Figure 6.1 Paracetamol listing[16]

paracetamol



Latin	paracetamolum
French	paracétamol
Spanish	paracetamol
Russian	парацетамол
Arabic	باراسيتامول
Chinese	对乙酰氨基酚
Phonetic	paracetamol
Molecular formula	$C_8H_9NO_2$
ATC Codes	N02BE01

National commission(s)	Alternate name(s)
JAN	acetaminophen
JP	acetaminophen
USP	acetaminophen
DCIT	paracetamolo

[16] To note, the "phonetic" listing in the Figure is typically identical to the orthographic form, and does not present the pronunciation in the International Phonetic Alphabet

Within the date range used for the dataset, there were 9,333 Russian names published up to October 2019 (the inclusion criteria for the corpus are in Section 5.2.1). The trilingual corpus includes names in three languages, for which the English form is for medications that are authorised in the UK (N=1,046 names). It is worth noting here that different medications are authorised in Russia than those authorised in the UK, so it may be that here there is some discussion of names for medications that are not in use in Russia, or medications that are authorised in Russia but not in the UK are missed in the discussion. Table 6.2 presents frequently prescribed medications and the names of their Russian counterparts.

Table 6.2 Names of top 50 most frequently prescribed medications in the UK and their Russian counterparts

English INN	***Russian INN***	***Back transliteration[17]***	***Therapeutic indication***	***Items prescribed annually in UK (millions)***
aciclovir	ацикловир	aciklovir	Antiviral used to treat herpes simplex infections	31.73
aflibercept	афлиберцепт	aflibercept	Used to treat neovascular (wet) age-related macular degeneration and metastatic colorectal cancer	59.69
amlodipine	амлодипин	amlodipin	Calcium channel blocker used to treat high blood pressure, coronary artery disease and variant angina	44.55
amoxicillin	амоксициллин	amoksicillin	Antibiotic used to treat bacterial infections such as pneumonia	53.72
apixaban	апиксабан	apiksaban	Anticoagulant used to prevent blood clots in conditions such as atrial fibrillation and deep vein thrombosis	25.02
atorvastatin	аторвастатин	atorvastatin	Statin used to prevent cardiovascular disease	67.36
beclometasone	беклометазон	beklometazon	Corticosteroid used to treat asthma and allergic rhinitis	64.80
benzydamine	бензидамин	benzidamin	Nonsteroidal anti-inflammatory drug (NSAID) used to relieve pain and inflammation, often in conditions like sore throat or mouth ulcers	63.57
bisoprolol	бисопролол	bisoprolol	Beta blocker used for heart diseases	51.89
bupivacaine	бупивакаин	bupivakain	Local anaesthetic	58.72
carbocisteine	карбоцистеин	karbocistein	Used to treat the symptoms of chronic obstructive pulmonary disease (COPD)	153.45

[17] Using ISO 9:1995 transliteration standard

carboplatin	карбоплатин	karboplatin	Chemotherapy medication used to treat various cancers, including ovarian, lung, and head and neck cancers	24.98
chlorhexidine	хлоргексидин	hlorgeksidin	Antiseptic used to clean and disinfect skin and mucous membranes, often used in oral care products	277.49
ciprofloxacin	ципрофлоксацин	ciprofloksacin	Antibiotic used to treat bacterial infections such as urinary tract infections (UTIs)	97.45
clozapine	клозапин	klozapin	Antipsychotic used to treat schizophrenia	50.29
colecalciferol	колекальциферол	kolekalciferol	A Vitamin D dietary supplement	72.20
desogestrel	дезогестрел	dezogestrel	Progestogen, used for birth control and treatment of menopausal symptoms	57.61
dexamethasone	дексаметазон	deksametazon	Corticosteroid with anti-inflammatory and immunosuppressive effects, used in various conditions such as asthma, allergies, and autoimmune diseases	38.99
fentanyl	фентанил	fentanil	Opioid used as an analgesic and sedative	72.88
flucloxacillin	флуклоксациллин	flukloksacillin	Antibiotic used to treat various infections, including skin infections	45.24
fluticasone	флутиказон	flutikazon	Corticosteroid used to treat asthma and allergic rhinitis	31.64
formoterol	формотерол	formoterol	Long-acting beta adrenoceptor agonist used in combination with corticosteroids to treat asthma and chronic obstructive pulmonary disease (COPD)	27.85
furosemide	фуросемид	furosemid	Loop diuretic used to treat oedema due to heart failure, liver scarring, or kidney disease	48.06
gabapentin	габапентин	gabapentin	Anticonvulsant medication used to manage seizures and neuropathic pain	26.58
ibuprofen	ибупрофен	ibuprofen	Nonsteroidal anti-inflammatory drug (NSAID) that is used to relieve pain, fever, and inflammation	64.14
iomeprol	йомепрол	jomeprol	Radiographic contrast agent used for diagnostic imaging procedures	25.45
iopamidol	йопамидол	jopamidol	Radiographic contrast agent used for diagnostic imaging procedures	34.93

lactulose	лактулоза	laktuloza	Laxative used to treat constipation and hepatic encephalopathy	299.24
lansoprazole	лансопразол	lansoprazol	Proton pump inhibitor used to reduce stomach acid	54.84
levetiracetam	леветирацетам	levetiracetam	Anticonvulsant medication used to treat seizures	25.24
levobupivacaine	левобупивакаин	levobupivakain	Local anaesthetic	68.76
levofloxacin	левофлоксацин	levofloksacin	Antibiotic used to treat a wide range of bacterial infections, including respiratory tract infections and urinary tract infections	
lidocaine	лидокаин	lidokain	Local anaesthetic	85.60
linezolid	линезолид	linezolid	Antibiotic used to treat infections caused by certain bacteria, including skin and soft tissue infections and pneumonia	36.08
metformin	метформин	metformin	Used to treat Type 2 diabetes	47.79
metronidazole	метронидазол	metronidazol	Antibiotic and antiprotozoal used either alone or with other antibiotics to treat pelvic inflammatory disease, endocarditis, and bacterial vaginosis	359.35
macrogol	макрогол	makrogol	Osmotic laxative used to relieve constipation	38.11
omeprazole	омепразол	omeprazol	Used to treat of gastroesophageal reflux disease (GERD), peptic ulcer disease, and Zollinger–Ellison syndrome	71.23
oxycodone	оксикодон	oksikodon	Opioid analgesic used to manage moderate to severe pain	41.39
paracetamol	парацетамол	paracetamol	Analgesic and antipyretic used to relieve pain and reduce fever	1,186.60
phenoxymethylpenicillin	феноксиметилпенициллин	fenoksimetilpenicillin	Antibiotic used to treat bacterial infections such as strep throat and tonsillitis	26.02
prednisolone	преднизолон	prednizolon	Corticosteroid, a steroid hormone used to treat certain types of allergies, inflammatory conditions, autoimmune disorders, and cancers	89.69

propofol	пропофол	propofol	Anaesthetic agent used for induction and maintenance of anaesthesia during surgery or sedation for procedures	174.95
ramipril	рамиприл	ramipril	ACE inhibitor type medication used to treat high blood pressure, heart failure, and diabetic kidney disease	42.13
ranitidine	ранитидин	ranitidin	H2 blocker used to reduce stomach acid production, primarily in the treatment of gastroesophageal reflux disease (GERD) and peptic ulcers	35.25
salbutamol	сальбутамол	salbutamol	Short-acting beta agonist used to relieve bronchospasm in conditions like asthma and COPD	222.20
sertraline	сертралин	sertralin	Selective serotonin reuptake inhibitor (SSRI) antidepressant used to treat depression, anxiety disorders, and obsessive-compulsive disorder (OCD)	25.82
sevoflurane	севофлуран	sevofluran	General anaesthetic	53.64
simvastatin	симвастатин	simvastatin	Statin medication used to lower cholesterol levels in the management of hyperlipidemia and reduce the risk of cardiovascular events	30.79
tacrolimus	такролимус	takrolimus	Immunosuppressant	45.72

6.3.2 Linguistic features of Russian-language INNs

Russian is a Slavic language, part of the broader Indo-European group, and in the same language family as Ukrainian and Bulgarian. Russian language INNs will be described here using the framework used in Chapter 4, exploring orthographic, morphosemantic, and relational features in turn. Selection of features to analyse in the two case studies is described in Section 6.2.1, with limitations.

6.3.2.1 Orthographic features of Russian-language INNs

I explored the character length and Scrabble score of all Russian-language INNs (measures Orth-01 and Orth-04). Russian languages contain a mean of 10.23 letters (SD 1.81). The shortest name has five letters (месна < mesna) and the longest name has 22 letters (феноксиметилпенициллин < phenoxymethylpenicillin). This is a little lower than English names (mean 10.60, SD 1.87). The Scrabble score in Russian is a little higher than in English (mean 1.94 compared with 1.72) but this may be due to a difference in scoring between the languages. Examples of modal names in Russian are ритуксимаб < rituximab; циталопрам < citalopram. Although the Russian alphabet has 33 letters, only 26 are present in the names in the dataset (ж, ч, ш, щ, ы, я, ъ are not used), so Russian orthography in the dataset is the same size as in English - 26 letters. Results are presented in Table 6.3.

Table 6.3 Orthographic features of Russian- and English-language INNs: length and Scrabble score

		Number of characters		***Scrabble score***	
		Russian INNs	*English INNs*	*Russian INNs*	*English INNs*
N	Valid	1046	1046	1046	1046
	Missing	0	0	0	0
Mean		10.24	10.60	1.94	1.72
Median		10.00	10.00	1.82	1.63
Std. Deviation		1.81	1.87	0.43	0.43
Range		17.00	18.00	2.53	2.50
Minimum		5.00	5.00	1.17	1.00
Maximum		22.00	23.00	3.70	3.50

Overall, orthographic features are very similar, pointing to a one-to-one grapheme exchange in transliteration from Latin into Cyrillic. Using the ISO 9:1995 transliteration standard, transliteration into Cyrillic from Latin is explored. In its basic form, the standard is a conversion table between Cyrillic and Latin letters: for example, г < > g; д < > d; н < > n. To test compliance with this standard in the dataset, Russian names are transliterated using the conversion table (Cubberley 2002:96, Table 9) and compared these back-formations with the English-language INNs with a distance measure (Levenshtein edit distance). This was used to quantitatively explore compliance with the transliteration standard.

Measuring distance highlights the transliteration mechanisms at play between Latin and Cyrillic. A common difference is that the final <e> at the end of English names is not transliterated, e.g. cabergoline > каберголин; cysteine > цистеин. Over half of English names end in a vowel (58%), predominantly <e> (44.4%, 464/1,046 names) and most names ending in an <e> lose the letter in the transliteration. This <e> is silent according to English phonology, such as in fludroxycortide and sulfasalazine, or in *write* and *bone* in the general language. However, though it is not syllabic, the presence of terminal <e> changes the pronunciation of the preceding vowel, lengthening it. Transliteration of Russian names is therefore driven by phonology, at least in part. Dropping the final <e> accounts for the slightly shorter name length of Russian names. This may work to correct unnecessary variation in the English nomenclature: for example, the STEM -ifene has two variant forms in English that do not differ in meaning (clomifene and tamoxIFEN) but this difference is lost in Russian (кломифен; тамоксифен) because the final <e> is dropped.

Endings are different between Russian and English INNs, but the transliteration approach is internally consistent. All English names ending in a consonant (577) are transliterated into a Russian name ending in a consonant (e.g. flucloxacillin > флуклоксациллин; acebutolol > ацебутолол). English names ending in a vowel (469) mostly lose that vowel if it is an <e> (447/469), e.g. spironolactone > спиронолактон, but keep it if it is an <a>, e.g. carbidopa > карбидопа. This is phonemically driven: a final <e> is not pronounced in English whereas a final <a> is pronounced. A handful transliterate the final <e> into a non-corresponding vowel (17/469). These are all enzymes with the STEM and chemical suffix -ase > -аза (e.g. streptokinase > стрептокиназа) or sugars with the chemical suffix -ose > -оза (e.g. lactulose > лактулоза), presumably because such terms exist independently of the INN nomenclature. Consequently, most Russian-language INNs in the dataset end in a consonant (97.7%), compared with only 55.2% of English-language INNs. The most common ending in Russian names is -ин, since a frequent ending in English is -ine (in a fifth of English names, 21.8%) and -in (in 12.2%), becoming -ин (in a third of Russian names 33.9%).

Although a letter is dropped, because this letter is silent this will not have an impact on the syllabic value of Russian names, because the number of syllables does not change. It is difficult to surmise the stress pattern of names in Russian, and whether they differ from that of English. Stress position varies in Russian, with some words stressed on the first syllable (e.g. '*солнце*, 'sun'), a middle syllable (e.g. *боль'ница*, 'hospital'), or with final stress (e.g. *моло'ко*, 'milk'). Stress can be fixed or mobile, and is affected by the speaker's familiarity with the word, its native status (whether it is a borrowing), and the case and number of the utterance, in terms of nouns (M. Shapiro, 1986).

In terms of pronunciation, Russian boasts a diverse range of sounds, including complex consonant clusters (e.g. *здравствуйте*, 'zdravstvuyte', 'hello'), palatalisation (e.g. the first /т/ in *телефон*, 'telephone', is palatalised), and vowel reduction. Vowel reduction is a prominent feature in Russian

phonology. Unstressed vowels are commonly reduced, such as in молоко, 'milk', which is stressed on the last syllable. The last <o> is pronounced as /o/, but the first and second <o> are reduced and centralised, to a sound similar to English schwa /ə/. This is also a common pronunciation feature in English, e.g. *sofa* /ˈsoʊ.fə/, in which the final unstressed <a> is reduced to a schwa /ə/ or /ɐ/. The orthography internal to the WHO nomenclature is smaller than the orthography of general languages - it has a close correspondence between letters and sounds (graphemes and phonemes). This is reflected in the close transliteration into Russian, because most phonemes in the English name have a corresponding Russian phoneme.

Excluding the difference in endings with a silent <e>, the majority of Russian names have a one-to-one grapheme correspondence with their English counterparts, according to the ISO conversion table, e.g. ацеклофенак < aceclofenac; лопразолам < loprazolam; метформин < metformin. Some letters vary by which letter they are transliterated by, such as <s>, which is variably transliterated as <з> and <с>, even within a single name (e.g. isosorbide > изосорбид). This seems to be driven somewhat by phonology, with the unvoiced <s> in аторвастатин < atorvastatin transliterated with unvoiced <с> and the (often) voiced <s> in преднизолон < prednisolone transliterated as voiced <з>. Some English letters do not have a Russian equivalent that is a single letter, so are transliterated by two letters, e.g. x > ks, e.g. amoxicillin > амоксициллин, acipimox > аципимокс. Other consonant clusters such as <qu> in English are transliterated variably as Russian <х>, e.g. atovaquone > атовахон, or Russian <кв>, e.g. bedaquiline > бедаквилин.

Russian is an inflected language and is case-determinate. This means it uses cases to determine the role of words in sentences, with six cases each having its own word endings (e.g. *друг*, 'friend' in nominative case; *с другом*, 'with a friend' in instrumental case). Russian words also change their forms to convey information like gender, number, and tense. Nouns are inflected for number and case, adjectives and pronouns also for gender. Nouns in Russian have gender distinctions (masculine, feminine, and neuter), requiring that related words like adjectives and pronouns match in terms of gender, number, and case. As nouns, drug names are subject to declension rules, e.g. *передозировка парацетамола* ('overdose of paracetamol', (ASNA, 2023). They may also be adjectivised, as in *парацетамольная токсичность* ('paracetamol toxicity'), though this process is likely to be names with which users are familiar and are used in the general language more frequently (Shapiro, 1986). Nouns ending in a consonant (in their singular nominative form) are treated as masculine in Russian, whereas nouns ending in an <a> are treated as feminine and nouns ending in an <o> as neuter (Cubberley, 2002). Most Russian names are likely treated as masculine because they end in a consonant.

How are letters prohibited in English treated in transl(iter)ation?

Several letters and letter clusters are prohibited for use in INNs by the WHO naming guidelines. These include: <y>, <h>, <th>, <ph>, <oe>, <ae>, <k>; and any duplicate letters such as <ll>. As described in full in Chapter 4, 192 names in English contained at least one prohibited letter or letter cluster. Prohibiting these letters is designed to facilitate international use of the names (as many of them do not exist in the alphabets of other languages, such as Italian and Spanish) and clusters such as <ph> can be supplanted by a single letter (<f>) rather than the traditional transcription of Greek <φ>. Chapter 4 found that use of these prohibited letters was a historic issue that is most prevalent in names approved in the 1950s and 1960s and not a continuing problem. Of course these names still persist in the system and refer to some of the most frequently prescribed names such as flucloxacillin and epinephrine. Furthermore, many of the prohibited letters are in the pharmacological STEM, so their use is unavoidable (e.g. <y> in -mycin and <ll> in -cillin).

I will discuss here in turn how each prohibited letter is treated in transliteration. Duplicate letters are exclusively transliterated into duplicates in Russian, e.g. amoxicillin > амоксициллин and argipressin > аргипрессин. The letter <y> is transliterated phonemically as <и>, e.g. capreomycin > капреомицин, with STEM -mycin. Transliteration works at a morphemic level where appropriate - <h>, when in the morph *chlor*, is transliterated as Russian *хлор*, e.g. chlorambucil > хлорамбуцил. When in the morph *meth* it is transliterated as *мет*, e.g. methyldopa > метилдопа. When in the morph *phos*, which has latterly been replaced by *fos* in English INNs, it is transliterated as *фос*, e.g. cyclophosphamide > циклофосфамид. The letter cluster <oe> is transliterated in full because in both cases in this dataset, it is not used to represent a single sound as in *oedema*, but rather sits on the boundary of two syllables, in moexipril > моэксиприл and pseudoephedrine >псевдоэфедрин. <ae> does not appear in the dataset. <k> is transliterated as Russian <k>.

With regards to Russian, the presence of prohibited letters in English does not necessarily present a problem. For example, Russian uses <к> for the hard /k/ sound and <с> for the soft /s/ sound (and also ц in some names, e.g. phenoxymethylpenicillin > феноксиметилпенициллин) - under WHO guidelines both of these sounds should be represented by a <c> in English, which can be ambiguous for pronunciation. Using <k> for the hard /k/ in English removes this ambiguity. For example, if ixekizumab was spelled as *ixecizumab (as the WHO guidelines indicate it should be), then it would be transliterated as *иксецизумаб, rather than the official иксекизумаб, and would change the pronunciation of the name. However, in many languages such as Italian and Spanish, the <k> is only present in foreign borrowings so would not be readily understood by all users. Use of <y> in English is unambiguous phonologically so readily transliterated into Russian <и>, as is <ph> (transliterated as <ф> in Russian). However, these are unnecessary in English and the smaller and shallow orthography recommended by the WHO would serve the transliteration process better. The phoneme /ð/ represented by the voiced <th> in English is not present in the Russian language so is supplanted by a <т> in all cases, as in

methyldopa > метилдопа. This provides some evidence of a consideration of morphology in the transliteration process.

6.3.2.2 Morphosemantic features of Russian-language INNs

Morphosemantic features of names in English were presented in Chapter 4, i.e. the way in which meaning is communicated using word parts, in this case STEMS. However, this was only made possible by having information on official STEMS from the WHO Stem Book. This information is not available for nomenclatures in languages other than English so instead the case studies in this chapter will provide a qualitative description of how STEMS are translated and transliterated, treating English as the 'source language' and Russian (and Italian in the next case study) as the 'target language' and surmising the STEM in translation. Stems were identified by looking at how the English STEM was transliterated, as the transliteration process on the whole was compliant with international guidelines (ISO 9 1995). They are not official or defined STEMS, but rather 'assumed' or 'observed' since they are the equivalent forms of English names that contain STEMS.

There are 197 orthographically-distinct STEMS in English names in the dataset used for this study, in a total of 1046 names. In Russian there are 196 orthographically-distinct STEMS, for 1046 names, because two English STEMS are transliterated into a single STEM in Russian (mycin, micin > мицин). The full list of STEMS is in the Appendix (A15) and the frequent STEMS are presented in A16, along with examples of names using this STEM, their corresponding Russian STEM, and the pharmacological group.

In 96.1% (744/774) of these names, the STEM in English is transliterated into Russian.[18] They generally seem to follow the ISO standard but are motivated by the phonology of the English source name, rather than orthography. As discussed above, STEMS with a final <e> that is not syllabic in English lose this in transliteration, e.g. clozapine > клозапин (not *клозапине). The position of the STEM does not change in translation into Russian in any name in the dataset, even in FREEFIXES where the position of the STEM is not fixed in the name according to WHO guidance, such as vir > вир in aciclovir > ацикловир. For example, lidocaine > лидокаин, tetracaine > тетракаин; and ceftriaxone > цефтриаксон.

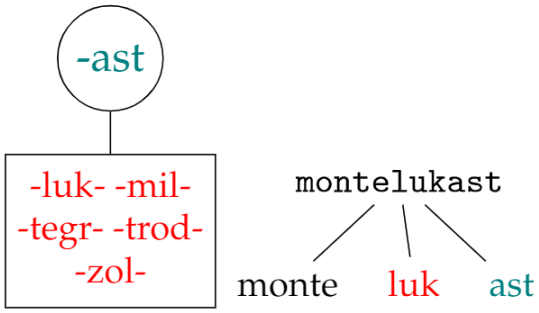
Headedness and treatment of SUBSTEMS

As outlined in Chapter 4, most SUBSTEMS are added to the existing STEM by agglutination, e.g. in monte[luk](ast) with 'parent' STEM -ast and 'child' SUBSTEM -luk-. By reading right to left, the user can first parse the name as an antiasthmatic (with the '-ast' STEM), then further sub-categorise as a leukotriene receptor antagonist (with the '-luk-' SUBSTEM). By agglutinating STEMS in this way, names show a clear

[18] 774 names in the dataset analysed as part of Chapter 4 are listed in the WHO Stem Book and contain a STEM compliant with WHO naming guidelines.

hyperonymic relationship between ‘parent’ and ‘child’ concepts and users can work right to left to decipher meaning. Figure 6.2 is reproduced here from Chapter 2. These names are right-headed, because they show their semantic head at the end of the name, and they make up around 69.0% of the subset of names analysed for headedness (312/452 names). This is replicated in Russian because names are transliterated so closely, e.g. монте[лук](аст).

Figure 6.2 Hyperonymic STEM building in montelukast (reproduced from Figure 2.3)



Around a third of names in English are not right-headed (~30%). In all cases, the position of the STEM in English is the same in Russian. A handful put their STEM at the beginning of the name (‘left-headed’), such as ceftriaxone > цефтриаксон , with cef > цеф. Others are FREEFIX STEMS, which can appear anywhere in the name without changing its meaning or the STEM taxon to which it belongs. For example, sita(gli)ptin > сита(гли)птин, with gli > гли. Their SUBSTEMS attach in various ways, even as a wrapper on either side of the STEM, as in ator[va](stat)[in] > атор[ва](стат)[ин], with parent STEM stat and SUBSTEM -vastatin. Because these names are not right-headed, they remain endocentric as they display their semantic head, but they are difficult to parse because their taxonomy is not displayed in a predictable order.

A small number of SUBSTEMS are not agglutinated to their STEM, but rather replace it (17 names)[19]. For example, lami[vudine] > лами[вудин] (replacing parent STEM -uridine), and sot[alol] > соталол (replacing parent STEM -olol). This replacement may involve a change in position (e.g. [ni]car[dipine] > [ни]кар[дипин] (replacing parent STEM nitro-). In one case, this involves the introduction of a negative SUBSTEM (e.g. tegafur > тегафур tegafur, replacing parent STEM -uridine). Consequently, these compounds no longer contain a head but rather a hyponym of the head in the form of the SUBSTEM, causing them to adopt an exocentric neoclassical compound structure. In all of these cases, the order of STEMS and SUBSTEMS is maintained in Russian translation, indicating that names are treated as full units and not as multi-componential terms.

[19] Replacement: 3 uridine; 1 pred-olone; 2 orphan; 2 olol; 3 drine; 4 azepam; 2 ster

In all cases, FREEFIX STEMS (e.g. calci, gab, gli, prost, vir) do not change position in translation and the name is transliterated as a standalone unit, rather than a multi-componential term. For example, the FREEFIX STEM vir > вир in aciclovir > ацикловир, enfuvirtide > энфувиртид, maraviroc > маравирок. A central question is the extent to which word formation processes can be observed in Russian, beyond transliteration (Panocová and ten Hacken 2022). There is some evidence of using word formation processes in translation into Russian, for example naming of enzymes (13 names, e.g. idursulfase > идурсульфаза) and sugars (five names, e.g. lactulose > лактулоза), as well as the form of STEMS being internally consistent to a taxon (observed in 96% of STEMS).

Motivation of STEMS and the impact of transl(iter)ation

In Chapter 4 the *motivation* of STEMS is explored, i.e. how STEMS are formed. STEMS are typically motivated by one of four mechanisms, though these are asserted based on first-hand observation of etymological links, rather than citing any published guidance or literature. The mechanisms are:

1. Abbreviating the name of the pharmacological group;
2. Forming a partial acronym of the name of the pharmacological group;
3. Abbreviating the chemical name; and
4. Abbreviating an existing INN.

In transl(iter)ation these links may be broken and this will lead to a loss in semantic motivation of the STEM. Each mechanism is discussed in turn, drawing examples from Appendix section A16 of frequent STEMS in English and Russian.

Type 1: Abbreviating the name of the pharmacological group

The first mechanism is by (ostensibly) abbreviating part of the name for the pharmacological group[20]. For example, -dil for vasodilators (carvedilol > карведилол), stat for enzyme inhibitors (also known as statins, e.g. atorvastatin > аторвастатин), and prost for prostaglandins (e.g. misoprostol > мизопростол). In translation, some of these STEMS may be decoupled from their semantic motivation in English, because the name for the pharmaceutical group will be different in Russian. It depends on the extent to which Russian uses neoclassical elements in pharmaceutical terms, because pharmaceutical language heavily uses neoclassical elements, e.g. vasodilators, with the combining form *vaso-* from Latin *vas,* vessel'.

I explored the translation of some of the pharmacological groups that are used to form the STEM, and found that many of them share etymological roots with the English term. Many of the Russian terms

[20] 107/452 names in frequent STEM taxa

describing the pharmacological group contain the translated STEM, so they replicate the semantic motivation when translated, with examples in Table 6.4. With others, there is not a direct translation but English and Russian share an etymological root. For example, -stim is contained in *colony stimulating factors* and there is an equivalent Russian verb for *stimulate* - ***стимулировать***, 'stimulirovat' that contains the Russian STEM, ***стим.*** In this case, a semantic link can be maintained in translation, though it may not be equivalent. Of course, it is difficult to know the extent to which users of medication names in any language understand technical terms and neoclassical forms, because the expertise of users is so wide-ranging (from patients to consultants). At a minimum at least the semantic motivation of these STEMS is maintained in translation to some extent and they are not completely arbitrary forms to Russian users.

Table 6.4 Examples of STEMS in English and Russian that are formed by abbreviating the name of the pharmacological group

English STEM	***Pharmacological group (English)***	***Russian STEM***	***Russian translation of pharmacological group, with pronunciation***
dil	vasodilators	дил	вазодилататоры (Pron. 'vazodilatori')
gab	gabamimetic agents	габ	габамиметические агенты (Pron. 'gabamimeticheskiye agenti')
gest	proGESTogens	гест	прогестагены (Pron. 'progestageny')
gli	(anti)hyperglycaemics	гли	гипергликемический (Pron. 'giperglikemicheskiy')
prost	prostaglandins	прост	простагландины (Pron. 'prostaglandiny')
sal	salicylic acid derivatives	сал	салициловая кислота (Pron. 'salitsilovaya kislota')
stat	statins, enzyme inhibitors	стат	статины, (Pron. 'statiny')
vir	antivirals	вир	антивирусные препараты (Pron. 'antivirusnye preparaty')

Type 2: Forming a partial acronym of the name of the pharmacological group

The second mechanism is to use the name of the pharmacological group to form an acronym[21]. These may use the first parts of each word or combining form inside the word, such as -mab for Monoclonal AntiBodies (e.g. infliximab > инфликсимаб; trastuzumab > трастузумаб). They also combine abbreviations and single letters, such as -imus for IMmUnoSuppressants (tacrolimus > такролимус; everolimus > эверолимус) and -rafenib for RAF (Rapidly Accelerated Fibrosarcoma) kinasE iNhIBitors; -tinib for Tyrosine kinase INhIBitors.

[21] 102/452 of names in frequent STEM taxa

I explored the translation of these STEMS in Russian: -imus; -mab; -rafenib; and -tinib. None of the partial acronyms in English can be formed in the same way in Russian, for various reasons. These include: Russian sentence structure differs from that of English; Russian agglutinates some adverbs and adjectives; and some Russian terms in pharmacology use combining forms that are not from Latin or Greek. *Immunosuppressants* in Russian is *иммунодепрессанты* (Pron. 'immunodepressanti'), which cannot be used in the equivalent way to form the acronym -imus, because the Russian form uses *депресс* rather than suppress. 'Monoclonal antibodies' in Russian is *моноклональные антитела* (Pron. 'monoklonalnyye antityela'), using *тела* for 'body', rather than a Latinised combining form such as *corpus*. Of course, the English form also uses a native (Germanic) form, *body*. This type of morphological calque can be found commonly in Russian, as well as in English and German (Puato, 2011). The STEM -маб will not hold the semantic link because the Russian term cannot be used to form that acronym, and instead would form the acronym *мат. *Raf (rapidly accelerated fibrosarcoma) kinase inhibitors* in Russian is *ингибиторы киназы Раф (быстрорастущий фибросаркома)*, pronounced as 'Ingibitory kinazy Raf (bystrorastushchiy fibrosarkoma)'. Russian reverses the word order to 'inhibitors of …', meaning that the acronym cannot be derived in the same way as in English. Furthermore, Russian agglutinates 'rapidly accelerating' into a single word *быстрорастущий*, 'big growing', not using neoclassical combining forms. The Russian translation of 'tyrosine kinase inhibitors' is *ингибиторы тирозинкиназ* (Pron. 'ingibitori tirozikinaz', lit. 'inhibitors of tyrosine kinase') which again reverses the word order, meaning that the acronym cannot be derived as in English.

Type 3: Abbreviating the chemical name

The third STEM formation mechanism is to abbreviate the chemical name, e.g. -retin for retinol derivatives, in tretinoin > третиноин.[22] As in the first mechanism of abbreviating the pharmacological group, the semantic motivation is likely maintained in translation as long as Russian uses a neoclassical form for the chemical compound or active ingredient. For the frequent STEMS in Table A8.3, most of the chemical names use the same neoclassical combining form in translation, so the semantic motivation of the STEM is maintained. In the case of mycin, which is derived from the genus name, the Russian translation also uses a Latinised form, *Стрептомицеты* ('Streptomitséty'), so the semantic link is maintained. Examples are presented in Table 6.5.

Table 6.5 Examples of STEMS in English and Russian that are formed by abbreviating the chemical name

English STEM	***Pharmacological group (English)***	***Chemical from which the STEM is derived***	***Russian STEM***	***Russian translation of chemical name, with pronunciation***
calci	vitamin D analogues/derivatives	calcium	кальци	кальций (Pron. 'kal'tsiy')

[22] 55/452 of names in frequent STEM taxa

-cillin	antibiotics, 6-aminopenicillanic acid derivatives	6-aminopenicillanic acid	циллин	6-аминопенициллановая кислота (Pron. '6-aminopenitsillanovaya kislota')
cef-	antibiotics, cefalosporanic acid derivatives	cefalosporanic acid	цеф	цефалоспорановая кислота (Pron. 'tsefalosporanovaya kislota')
-mycin	antibiotics, produced by Streptomyces strains (see also -kacin)	Streptomyces	мицин	Стрептомицеты (Pron. 'Streptomitséty')
-retin	retinol derivatives	retinol	ретин	ретинол (Pron. 'retinol')
-tide	peptides and glycopeptides (for special groups of peptides see -actide, -pressin, -relin, -tocin)	peptides, glycopeptides	тид	пептиды, гликопептиды, ('Pron 'peptidy, glikopeptidy')

Broadly speaking, neoclassical combining forms are used in the Russian language but some neoclassical combining forms have native equivalents, such as ***водоро́д*** ('vodorod') for hydrogen, which uses ***вода***, 'water'. This is in contrast with the Greek combining form гидро, such as in ***гидрология*** (hydrology, pron. 'gidrologiya'). This is common in languages not closely related to Latin or Ancient Greek, such as German and Dutch (Puato, 2011). The combining form *chlor* in chlorine is typically translated as *хлор* in Russian such as in ***хлороводоро́дная кислота*** (hydrochloric acid, pron. 'chlorovodorodnaya kislota'), and does not have a native equivalent. In the dataset, names using hydro are translated by the Greek form, e.g. diphenhydramine > дифенгидрамин; dihydrocodeine > дигидрокодеин; hydrochlorothiazide > гидрохлоротиазид; hydroxyzine > гидроксизин. In these cases combining forms are not treated as separate morphs. French was spoken by the Russian aristocracy in the 18th and 19th centuries and many neoclassical forms persist today in technical language, so Russian speakers are likely to understand the meanings of neoclassical forms, to varying extents.

One STEM is a standard chemical suffix: -ase, for enzymes. Russian names for enzymes end in аза, e.g. amylase > амилаза (pron. 'amilaza') and lactase > лактаза (pron. 'laktaza'). The Russian translations of names in the -ase taxon do not follow the transliteration method of other names with a final -e that is not syllabic, i.e. not pronounced. In these names, the STEM is treated as a morph and not transliterated phonemically. For example, streptokinase > стрептокиназа (pron. 'streptokinaza') and cocarboxylase > кокарбоксилаза (pron. 'kokarboxylaza').

Type 4: Abbreviating an existing INN

Many INNs are ostensibly named after the first INN in that STEM family (as the INN is mentioned in the definition of the pharmacological group)[23]. It is unclear in these cases what the motivation for that first INN was, and it may be that it is arbitrary. For example, -azepam is the STEM for diazepam derivatives (diazepam was approved in 1965) and -oxetine for fluoxetine derivatives (fluoxetine was approved in 1976). In these cases, the first INN in the taxon serves a semantic function and is the relative motivation for others in the taxon. This motivation will be maintained in translation because Russian INNs are close transliterations of English INNs, and the STEMS are derived from these close transliterations. For example: ketoconazole > кетоконазол; zolmitriptan > золмитриптан. In some cases the first INN has since been discontinued, e.g. ibufenac in the -ac taxon, for "anti-inflammatory agents, ibufenac derivatives", e.g. ketorolac > кеторолак; sulindac > сулиндак.

Unclear semantic motivation

In many cases the semantic motivation of the STEM is unclear, as in -pril for "ACE inhibitors", such as captopril > каптоприл; and -terol for "bronchodilators, phenethylamine derivatives", such as salmeterol > салметерол[24]. Since the Russian transliteration is so close to the English INNs, the STEM is in the same position and form in translation, but semantic motivation is just as unclear.

6.3.2.3 Relational features of Russian-language INNs

This section looks at relational features of Russian-language INNs, i.e. how names interact with other names in the dataset. In Chapter 4 ten linguistic features of INNs that are relational are explored. Because STEM information is not available for Russian INNs, five of the ten measures are possible in this language: the frequency of two-letter and three-letter strings at the beginnings and endings of names (so-called *ngrams*) and neighbourhood density. Neighbourhood density is the number of names within a fixed distance of a target name. In Chapter 4, neighbours were identified by measuring the similarity between names in the dataset and defining a neighbour as any name sharing more than 30% of three-letter strings. Two spaces were added to the beginning of names to weight the beginning, as in Lambert et al. (1999). For example the neighbourhood of fluconazole includes: flucytosine, ketoconazole, miconazole and eight other names. Names sharing their first letter have been linked with LASA error (in Shah et al., 2017), as have names with a high neighbourhood density (Lambert et al., 2001b).

Beginnings and endings of Russian INNs

As in the English-language INN analysis in Chapter 4, frequency of two-letter strings (bigrams) and three-letter strings (trigrams) are explored at the beginnings of Russian names and the results compared

[23] 83/452 of names in frequent STEM taxa

[24] 92/452 of names in frequent STEM taxa

between English and Russian. The most frequent (top ten) initial and terminal trigrams in Russian and English INNs are given in Table 6.6. Sections A17 to A20 in the Appendix give supplementary information on distribution and frequency of initial and terminal ngrams in Russian.

The top five beginnings are the same in each language (bigrams *ме, ка, пр, тр, те*; trigrams *про, флу, кар, мет, кло*). This backs up earlier observations of a close transliteration from English into Russian. The most frequent initial bigram of INNs in Russian is ме (e.g. мебеверин < mebeverine) and most frequent trigrams are флу and про (e.g. пропилтиоурацил < propylthiouracil). The distribution of initial strings is comparable between Russian and English, indicating very little difference in the way names are started in both languages (Russian names have 545 unique initial trigrams, and English names have 567). Names sharing the first letter and sharing strings of letters have been linked with LASA error (Shah et al., 2017) so if the distribution of initial strings is similar in Russian to that of English then the risk is relatively similar in Russian-language names.

The most frequent two-letter ending of INNs in Russian is *ин* (fludarabine > флударабин) and most frequent three-letter ending is *маб* (infliximab > инфликсимаб. This differs from the most frequent trigram ending in English, which is (by a long way) *ine*, appearing in more than a fifth of INNs in the dataset. As outlined in Chapter 4, endings of English names are more homogeneous than their beginnings, with more than a quarter ending in *ine* or *ide*. In English names there are more than 132 distinct terminal trigrams, compared with 567 distinct initial trigrams. By not transliterating the final <e> in these trigrams, the resulting endings become more diverse in Russian - Russian names contain 148 distinct terminal trigrams. For example, levobupivacaine (-ine) is transliterated into Russian as левобупивакаин (-ин), so the trigram string measured is -аин. In essence, a frequent trigram in English has become a frequent bigram in Russian. Two-letter endings are similarly homogeneous to English, however, with more than a third of names ending in ин.

Table 6.6 Initial and terminal trigrams in Russian and English (top ten most frequent, in descending order of frequency)

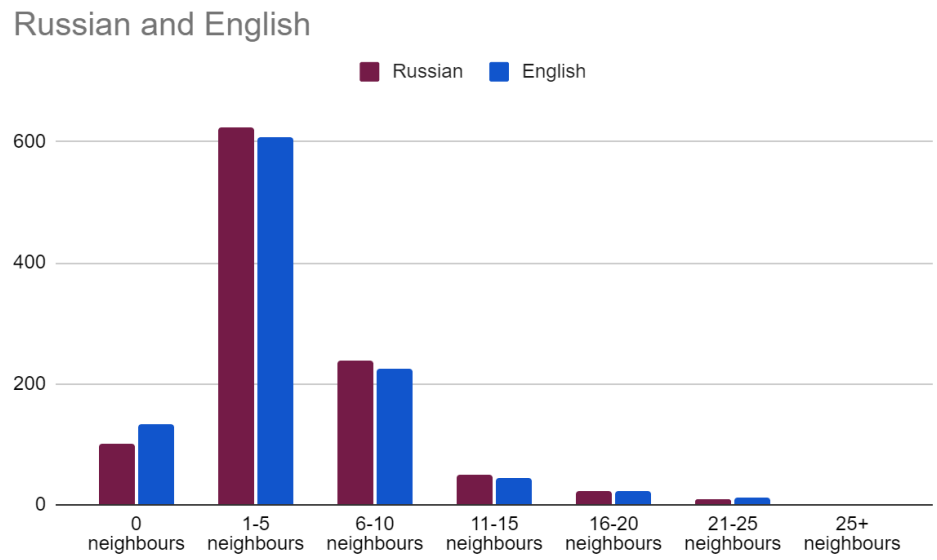
	Initial trigrams		***Terminal trigrams***	
Russian	флу- про- кар- мет- кло- три-	хло- лев- окс- фен- цеф-	-маб -тин -лин -цин -зин	-ниб -мид -дин -вир -мин
English	flu- pro- car- met- chl-	clo- tri- cef- lev- hyd-	-ine -ide -one -mab -nib	-cin -vir -ole -tin -tan

Neighbourhood density of Russian names

Neighbourhood density in Russian was similar to that of English. Results are compared between languages in Figure 6.3 and distribution is given in the Appendix section A21. The majority of names have between one and five neighbours (623/1046, 59.6%), as in English (607/1046, 58.0%). Russian neighbourhood density is closely correlated with that of English (Pearson rho=0.912, p<0.001). Many names have an identical similarity neighbourhood in Russian as in English, e.g. лансопразол < lansoprazole, both with the same seven orthographic neighbours, including арипипразол < aripiprazole and алпразолам < alprazolam, and севофлуран < sevoflurane with the same three orthographic neighbours, including десфлуран < desflurane and метоксифлуран < methoxyflurane.

Some differences in neighbourhood are due to a difference in word length, which skews the calculation used to calculate orthographic similarity (the Dice coefficient, which is the proportion of shared strings to total strings). For example, clozapine and clonazepam have a trigram-2b score of 31%, but клозапин and клоназепам have a score of 33% because клозапин is shorter than clozapine. Since the cut-off is 33%, in English these are not neighbours whereas in Russian there are. In other cases, the phonemically-driven orthography used in Russian compared with English means that Russian names share more strings so have more neighbours. For example, amoxicillin and phenoxymethylpenicillin are 29% similar, but амоксициллин and феноксиметилпенициллин are 41% similar, because English uses <y> instead of <i> in ‘oxy’. In this case, a smaller orthography and avoiding prohibited letters increases similarity in the nomenclature. Russian names for highly frequently prescribed medications all have at least one neighbour, but up to 12 neighbours (e.g. клозапин, with: триклозан, клобетазол, клобетазон, кармеллоза, кломифен, клонидин, клофарабин, клоназепам, клобазам, кветиапин, миртазапин, оланзапин).

Figure 6.3 Distribution of neighbourhood density in Russian



This concludes the case study of Russian-language INNs. The next case study in Section 6.4 looks at Italian common names. Findings from both case studies are discussed in Section 6.5.

6.4 Case study 2: Italian-language names (*Denominazioni Comuni Italiane*, *DCIt*)

6.4.1 Denominazioni Comuni Italiane (Italian generic names for medications): overview

This case study explores linguistic features of Italian generic names for medications. In Italy, national generic names for medications are *Denominazioni Comuni Italiane*, 'DCIt'. Where possible, linguistic features between Italian and English names are compared. This case study is the second in this chapter, following a case study of Russian-language international nonproprietary names (INNs). In the discussion section of this chapter findings are synthesised and linguistic features compared between Russian and Italian nomenclatures.

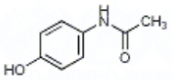
INNs are first published in English by the World Health Organization (WHO), then translated into the other five official languages of the United Nations, including Russian. They are also translated into Latin (SoINN, 2020b). The WHO does not issue any guidance on translation into other languages outside of these seven, but some sources state that the Latin form of the name is used as the source form for translation into other European languages, such as Italian and Portuguese (Marecková & Cerveny, 2002). Approving an INN for a medicinal product is part of the process of publishing the International Pharmacopoeia (Ph. Int.), a volume covering all aspects of pharmacology for approved medications. The Ph. Int. is published by the WHO and accessible online (WHO, 2023). The European Pharmacopoeia

(Ph. Eur.) is derived from the Ph. Int. and is published every three years by the Council of Europe in English and French, with names also published in Latin (Council of Europe, 2023a), with examples here: (Council of Europe, 2023b) Member states of the Ph. Eur. then publish a national pharmacopoeia in their national language(s). In the case of Italy, this is the *Farmacopea Ufficiale Italiana* (FUI, lit. 'official pharmacopoeia of Italy'), which includes a translation into Italian generic names, *Denominazioni Comuni Italiane* (DCIt, lit, 'Italian Common Names').

Since names in the Ph. Eur. are published in English, French, and Latin, it is unclear which language is used to derive the Italian name. It may be from the Latin name, given the close relationship between Italian and Latin. Latin names have not been collected as part of this study and here English and Italian names are compared. It is worth noting that while English is used as the *de facto* source language for translation into other UN languages and is analysed in this thesis and compared with Russian in the previous case study, in the case of Italian the source language may in fact be Latin. Figure 6.4 is reproduced from Figure 6.1and shows names in the UN languages, Latin, and Italian.

Figure 6.4 Paracetamol listing (reproduced from Figure 6.1)

paracetamol



Latin	paracetamolum
French	paracétamol
Spanish	paracetamol
Russian	парацетамол
Arabic	باراسيتامول
Chinese	对乙酰氨基酚
Phonetic	paracetamol
Molecular formula	$C_8H_9NO_2$
ATC Codes	N02BE01

National commission(s)	Alternate name(s)
JAN	acetaminophen
JP	acetaminophen
USP	acetaminophen
DCIT	paracetamolo

Within the date range used for the dataset (from beginning of the INN programme in 1951 to end of October 2019), 9,333 INNs were published. An inclusion criterion of the dataset is that the medication should be authorised in the UK, resulting in 1,046 names. These names formed the basis of the trilingual corpus in English, Russian, and Italian. 22 of these names could not be found in the online

pharmacopoeia, so the Italian corpus contains 1,024 names. Because of that inclusion criterion, in this case study it is possible that some medications are missing that are authorised in Italy but not in the UK. However, these medications are listed in the Italian Pharmacopeia so they are authorised in Italy. Table 6.7 presents frequently prescribed medications and the names for their Italian counterparts.

Table 6.7 Names of top 50 most frequently prescribed medications in the UK and their Italian counterparts

English INN	***Italian DCIt***	***Therapeutic indication***	***Items prescribed annually in UK (millions)***
aciclovir	aciclovir	Antiviral used to treat herpes simplex infections	31.73
aflibercept	aflibercept	Used to treat neovascular (wet) age-related macular degeneration and metastatic colorectal cancer	59.69
amlodipine	amlodipina	Calcium channel blocker used to treat high blood pressure, coronary artery disease and variant angina	44.55
amoxicillin	amoxicillina	Antibiotic used to treat bacterial infections such as pneumonia	53.72
apixaban	apixaban	Anticoagulant used to prevent blood clots in conditions such as atrial fibrillation and deep vein thrombosis	25.02
atorvastatin	atorvastatina	Statin used to prevent cardiovascular disease	67.36
beclometasone	beclometasone	Corticosteroid used to treat asthma and allergic rhinitis	64.80
benzydamine	benzidamina	Nonsteroidal anti-inflammatory drug (NSAID) used to relieve pain and inflammation, often in conditions like sore throat or mouth ulcers	63.57
bisoprolol	bisoprololo	Beta blocker used for heart diseases	51.89
bupivacaine	bupivacaina	Local anaesthetic	58.72
carbocisteine	carbocisteina	Used to treat the symptoms of coronary obstructive pulmonary disorder (COPD)	153.45
carboplatin	carboplatino	Chemotherapy medication used to treat various cancers, including ovarian, lung, and head and neck cancers	24.98
chlorhexidine	clorexidina	Antiseptic used to clean and disinfect skin and mucous membranes, often used in oral care products	277.49
ciprofloxacin	ciprofloxacina	Antibiotic used to treat bacterial infections such as urinary tract infections (UTIs)	97.45
clozapine	clozapina	Antipsychotic used to treat schizophrenia	50.29
colecalciferol	colecalciferolo	A Vitamin D dietary supplement	72.20
desogestrel	desogestrel	Used for birth control and treatment of menopausal symptoms	57.61

dexamethasone	desametasone	Corticosteroid used for its anti-inflammatory and immunosuppressive effects, employed in various conditions such as asthma, allergies, and autoimmune diseases	38.99
fentanyl	fentanil	Opioid used as an analgesic and sedative	72.88
flucloxacillin	flucloxacillina	Antibiotic used to treat various infections, including skin infections	45.24
fluticasone	fluticasone	Corticosteroid used to treat asthma and allergic rhinitis	31.64
formoterol	formoterolo	Long-acting beta agonist used in combination with corticosteroids to treat asthma and chronic obstructive pulmonary disease (COPD)	27.85
furosemide	furosemide	Loop diuretic medication used to treat edema due to heart failure, liver scarring, or kidney disease.	48.06
gabapentin	gabapentin	Anticonvulsant medication used to manage seizures and neuropathic pain	26.58
ibuprofen	ibuprofene	Nonsteroidal anti-inflammatory drug (NSAID) that is used to relieve pain, fever, and inflammation.	64.14
iomeprol	iomeprolo	Radiographic contrast agent used for diagnostic imaging procedures	25.45
iopamidol	iopamidolo	Radiographic contrast agent used for diagnostic imaging procedures	34.93
lactulose	lattulosio	Laxative used to treat constipation and hepatic encephalopathy	299.24
lansoprazole	lansoprazolo	Proton pump inhibitor used to reduce stomach acid	54.84
levetiracetam	levetiracetam	Anticonvulsant medication used to treat seizures	25.24
levobupivacaine	levobupivacaina	Local anaesthetic	68.76
levofloxacin	levofloxacina	Antibiotic used to treat a wide range of bacterial infections, including respiratory tract infections and urinary tract infections	
lidocaine	lidocaina	Local anaesthetic	85.60
linezolid	linezolid	Antibiotic used to treat infections caused by certain bacteria, including skin and soft tissue infections and pneumonia	36.08
metformin	metformina	Used to treat Type 2 diabetes	47.79
metronidazole	metronidazolo	Antibiotic and antiprotozoal used either alone or with other antibiotics to treat pelvic inflammatory disease, endocarditis, and bacterial vaginosis	359.35
macrogol	macrogol	Osmotic laxative used to relieve constipation	38.11

omeprazole	omeprazolo	Used to treat of gastroesophageal reflux disease (GERD), peptic ulcer disease, and Zollinger–Ellison syndrome	71.23
oxycodone	ossicodone	Opioid analgesic used to manage moderate to severe pain	41.39
paracetamol	paracetamolo	Analgesic and antipyretic used to relieve pain and reduce fever	1,186.60
phenoxymethylpenicillin	fenossimetilpenicillina	Antibiotic used to treat bacterial infections such as strep throat and tonsillitis	26.02
prednisolone	prednisolone	Corticosteroid, a steroid hormone used to treat certain types of allergies, inflammatory conditions, autoimmune disorders, and cancers	89.69
propofol	propofol	Anesthetic agent used for induction and maintenance of anesthesia during surgery or sedation for procedures	174.95
ramipril	ramipril	ACE inhibitor type medication used to treat high blood pressure, heart failure, and diabetic kidney disease	42.13
ranitidine	ranitidina	H2 blocker used to reduce stomach acid production, primarily in the treatment of gastroesophageal reflux disease (GERD) and peptic ulcers	35.25
salbutamol	salbutamolo	Short-acting beta agonist used to relieve bronchospasm in conditions like asthma and COPD	222.20
sertraline	sertralina	Selective serotonin reuptake inhibitor (SSRI) antidepressant used to treat depression, anxiety disorders, and obsessive-compulsive disorder (OCD)	25.82
sevoflurane	sevoflurano	General anaesthetic	53.64
simvastatin	simvastatina	Statin medication used to lower cholesterol levels in the management of hyperlipidemia and reduce the risk of cardiovascular events	30.79
tacrolimus	tacrolimus	Immunosuppressant	45.72

6.4.2 Linguistic features of Italian generic names

Italian is a Romance language, part of the broader Indo-European group, and in the same language branch as Spanish, French, and Romanian. Italian names will be described here using the framework used in Chapter 4, exploring orthographic, morphosemantic, and relational features in turn. Selection of features to analyse in the two case studies is described in Section 6.2 and Table 6.1, and includes consideration of limitations.

6.4.2.1 Orthographic features of Italian names

I explored the length and Scrabble score (normalised by length) of all Italian names in the dataset (n=1024). These are measures Orth-01 and Orth-04 in Chapter 4, and descriptive statistics are given in Table 6.8. Italian names contain a mean of 10.63 (SD 1.86) letters. The shortest name has five letters (mesna < mesna) and the longest name has 23 letters (fenossimetilpenicillina < phenoxymethylpenicillin). This is similar to English names (mean 10.60, SD 1.87). The Scrabble score is quite a bit higher in Italian (2.13 mean) but this is due to the method of scoring. Because <k>, <w>, and <x> are not letters in the Italian alphabet, these letters were given a score of 10 and consequently several names containing these letters have pushed up the score. Examples of modal names in Italian are velpatasvir < velpatasvir and felodipina < felodipine. All 21 letters of the Italian alphabet are used in Italian names in this dataset, plus the letters <k> (e.g. amikacina), <w> (e.g. warfarin) and <x> (e.g. xipamide) that are used in loanwords. However, these occur in small numbers.

Table 6.8 Orthographic features of Italian and English names: length and Scrabble score

		Number of characters		***Scrabble score***	
		Italian names	*English INNs*	*Italian names*	*English INNs*
N	Valid	1024	1046	1024	1046
	Missing	0	0	0	0
Mean		10.63	10.60	2.13	1.72
Median		10.00	10.00	2.09	1.63
Std. Deviation		1.86	1.87	0.49	0.43
Range		18.00	18.00	3.30	2.50
Minimum		5.00	5.00	1.00	1.00
Maximum		23.00	23.00	4.30	3.50

Almost half of English names do not change in the translation process to Italian, such as apixaban < apixaban and citalopram < citalopram (502/1024, 49.0%)[25]. When a name does change at the end, the name gains a vowel or substitutes one vowel for another. This can be by adding a vowel to a consonant; in this dataset, this only happens to names ending in <l>, <n>, or occasionally <d>. For example, English *ol* becomes *olo* in Italian: salbutamol > salbutamolo; and English *in* becomes *ina* in Italian: sertraline > sertralina. It can also be by substituting the non-syllabic final <e> in English names for a syllabic <a>, <o>, or <i> in Italian, such as in sulfadiazine > sulfadiazina. This <e> is silent according to English phonology, such as in fludroxycortide and sulfasalazine, or in *write* and *bone* in the general language. However, though it is not syllabic, the presence of terminal <e> changes the pronunciation of the preceding vowel, lengthening it. In Italian, a final <e> is pronounced, such as in *amore* (/aˈmoːre/), with three syllables. Because of this change in syllabic value, even names that do not undergo an orthographic change will still be subject to a phonemic change. For example, dapsone (/dapˌsɔnɛ/) < dapsone (/ˈdæpsoʊn/) gains a syllable. In some cases, unnecessary variation in English orthography is removed in translation into Italian, e.g. the STEM -ifene has a variant -ifen (clomifene and tamoxifen), but both are -ifene in Italian (e.g. clomifene and tamoxifene).

With the addition of vowels at the end of names, and the phonemic change of the final <e>, the majority of Italian names in this dataset have an additional syllable to their English counterpart. A change in syllabic value will affect stress of the name, and this can lead to secondary stress in longer names. Standard stress pattern in Italian is on the penultimate syllable, e.g. paracetamolo and omeprazolo (e.g. PharmaStarTV, 2015; Salute, 2020). However, many names are also stressed on the final syllable (*caffè*) or the prepenultimate syllable (*portico*).

Consequently, most Italian names in the dataset end in a vowel (63.61%, 665/1024), compared with only 44.83% (469/1046) of English names. The most common ending in Italian names is *-ina* (in 30.31%, 317/1024), since frequent endings in English are *-ine* (in a fifth of English names, 21.8%) and *-in* (in 12.2%). Italian nouns are assigned a gender, with regular feminine nouns ending in an <a> in the singular (e.g. *macchina*, 'car'), and regular

[25] A notable exception is epinephrine which is translated into Italian as adrenalina, following the European model of retaining its Latin-based name to avoid confusion errors for this emergency medication. In contrast, Russian transliterates the English name epinephrine, as эпинефрин (see Aronson, 2000).

masculine nouns ending in an <o> in the singular (e.g. *libro*, 'book'). Nouns ending in an <e> can be either feminine (e.g. *la notte,* 'the night') or masculine (*il cane*, 'the dog'). Nouns in Italian that end in a consonant are typically foreign loanwords in Italian and are masculine, e.g. *il film* and *il jazz*. In the Italian dataset, 336 names end in <a> and 129 end in <o> so 32.81% are treated as feminine and 12.60% are treated as masculine. In the remaining names, the gender cannot be determined from the form of the name.

The orthography internal to the English-language INN nomenclature is smaller than the orthography of the general language and consequently it has a close correspondence between letters and sounds (graphemes and phonemes). For example, English-language names do not contain word parts such as <eigh> as in *height* or <tch> as in *watch*, that are not aligned to the phonemes of individual letters. Consonants in names also undergo phonemic change, without an orthographic change. For example, in procarbazine > procarbazina, English <z> is voiced and becomes unvoiced in Italian, pronounced /ts/. Other consonant clusters such as <qu> in English become <qu>, e.g. atovaquone > atovaquone, or <ch>, e.g. chloroquine > clorochina, in Italian. Some consonant clusters that are not standard in Italian phonotactics are changed, e.g. <ct> > <tt> in lactulose > lattulosio and spironolactone > spironolattone (see *pact* > *patto* cognate in general language).

Around two thirds of double or triple vowel combinations are pronounced as individual syllables in English. Double and triple vowel combinations present a number of challenges. Firstly, they create ambiguity in pronunciation, with some double vowels pronounced as a single-syllable vowel sound (e.g. acenocOUmarol > acenocumarolo), a single-syllable diphthong (e.g. parathyrOId), or as two syllables (e.g. phenytOIn). This will lead to variant pronunciations across languages - for example, Italian is a highly phonological language and unusual double vowels (such as the EU in aldeskEUkin) would likely be pronounced as two syllables whereas in English with one. Since pronunciation of INNs (or indeed any medication names) is not proscribed, but rather assumed to be obvious, many names will likely have multiple pronunciations depending on language used and the regional variant of that language. This also reveals a limitation to this part of the analysis, namely that many of the pronunciations were decided subjectively by RB, and confirmed by watching videos online about the medications. This leaves considerable room for error, and for regional language variants.

<u>How are letters prohibited in English treated in translation?</u>

Several letters and letter clusters are prohibited for use in INNs by the WHO naming guidelines. These include: <y>, <h>, <th>, <ph>, <oe>, <ae>, <k>; and any duplicate letters such as <ll>. As described in full in Chapter 4, 192 names in English contained at least one prohibited letter or letter cluster. The rationale behind not permitting these letters in INNs is to facilitate translation and international use of the names (as many of them do not exist in the languages of other alphabets, such as Italian and Spanish) and clusters such as <ph> can be supplanted by a single letter (<f>) rather than the Greek form of the phoneme. Chapter 4 found that use of these prohibited letters was a historic issue most prevalent in names approved in the 1950s and 1960s and is not an enduring problem. Of course, these names still persist in the system and refer to some of the most frequently prescribed names such as flucloxacillin and epinephrine. Many of the prohibited letters are in the pharmacological STEM, so their use is unavoidable (e.g. <y> in -mycin and <ll> in -cillin). They also appear heavily in neoclassical combining forms for chemical substances, such as meth (from Ancient Greek *μέθυ*, 'wine'), chlor (from Ancient Greek *χλωρός*, 'pale green').

Although Italian names are not subject to this guideline (because they are not designated by the WHO), the prevalence of these letters in Italian names is explored, and how they are treated in translation from English to assess the usefulness of this guideline to facilitate translation. There are not many of these prohibited letters in the Italian dataset. The letter <y> is treated phonemically as <i>, such as in acetylcysteine > acetilcisteina, with STEM -steine. The (Ancient) Greek-derived <ph> is also treated phonemically as <f> in Italian, such as in chloramphenicol > cloramfenicolo. Morphs for chemical groups are typically translated on a morpheme- rather than letter-based level, such as *meth* > *met* (e.g. dexamethasone > desametasone); *chlor* > *clor* (e.g. chlorambucil > clorambucil); *oxy* > *ossi* (e.g. oxycodone > ossicodone); and *hydro* > *idro* (e.g. hydroxyzine > idroxizina). All duplicate letters in English are also duplicates in Italian, though it is worth noting that this is a relatively new rule stipulated by the WHO in 2020 so was not applicable when the majority of names in this dataset were created (SoINN, 2020a), e.g. cinnarizine > cinnarizina). The only prohibited letter to appear consistently in the Italian nomenclature is <k>, such as in amikacin > amikacina. Where the English name uses a <k>, so will the Italian name, unless the <k> forms part of a morph that has a chemical meaning. For example, streptokinase already has a translation in Italian: streptochinasi.

The presence of prohibited letters in English does not consistently present a problem for translation into Italian. Although <k> is not in the Italian alphabet, it is in foreign loanwords and widely understood (e.g. *koala* and *ketchup*). The hard /k/ sound is represented by either <c> (e.g. *cane*) or <ch> (*chiesa*). <k> is useful in English names to distinguish between hard

and soft sounds (/k/ vs /s/, in clozapine and amoxicillin) but is not consistently used because it is prohibited. However, in the smaller orthography of the WHO using <k> for the hard /k/ sound can actually *remove* ambiguity - e.g. if ixekizumab was spelled as *ixecizumab (as the WHO guidelines indicate it should be), then its pronunciation would be ambiguous. The letter clusters <ph> and <y> are unnecessary and have alternative letters that can be used in both English and Italian, so these present an unnecessary challenge in the translation process, though they are unambiguous phonemically. The sounds represented by <th> sound do not exist in standard Italian. By prohibiting the letter group in English they are also prohibiting the sound, but the morph meth is a well-established combining form in English. Because it is derived from Greek, it has neoclassical counterparts in Italian and so does not present a challenge in translation, as in methyldopa > metildopa. In short, some prohibited letters can remove ambiguity in translation, some are unnecessary, and some are acceptable when they form part of well-established neoclassical combining forms such as *meth* and *chlor*.

6.4.2.2 Morphosemantic features of Italian names

Chapter 4 looks at morphosemantic features of names in English, i.e. the way in which meaning is communicated using STEMS. However, this was only made possible by having information on official STEMS from the WHO Stem Book (WHO, 2017a). This information is not available for nomenclatures in languages other than English so instead the case studies in this chapter will provide a qualitative description of how STEMS are translated and transliterated, treating English as the 'source language' and Italian (and Russian in the previous case study) as the 'target language' and surmising the STEM in translation. Italian STEMS were identified by looking at how the English STEM was translated. They are not official or defined STEMS, but rather 'assumed' or 'observed' since they are the equivalent forms of English names that contain STEMS.

There are 197 orthographically-distinct STEMS in English names in the dataset used for this study, in a total of 1046 names. In Italian there are 196 orthographically-distinct STEMS, in 1024 names, because two English STEMS are transliterated into a single STEM in Italian (mycin, micin > micina in Italian. The full list of STEMS is in the Appendix (section A15) and the frequent STEMS are presented in Appendix section A22, along with examples of names using this STEM, their corresponding Italian STEM, and the pharmacological group.

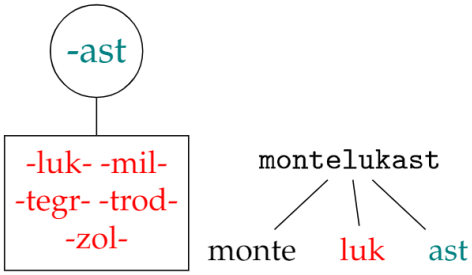
In 64% (487/761[26]) of names that contain an official STEM, the English STEM is the same as the Italian STEM, e.g. nitrazepam > nitrazepam. A further 35.3% (269/761) differ only by the last letter, typically Italian <a> supplanting <e>, as in acrivastine > acrivastina. The position of the STEM does not change in translation into Italian in any name in the dataset, even in FREEFIXES where the position of the STEM is not fixed in the name according to WHO guidance, such as vir > vir in aciclovir > aciclovir. All Italian STEMS are in the same position as the STEM in their English counterpart. STEMS are not treated as components in the name, but rather the name is treated as a single unit and transliterated in full. For example, lidocaine > lidocaina, tetracaine > tetracaine; and ceftriaxone > ceftriaxone. STEMS are likely to undergo a change in translation into Italian because most STEMS are SUFFIXES and many English names that end in a consonant gain a terminal vowel in English (e.g. cillin > cillina, astina > astina). STEMS will also change phonemically due to pronunciation differences between English and Italian, even when there is no orthographic change (e.g. erythromycin > eritromicina; /ɪˌrɪθrəˈmaɪsɪn/ > /eritromiˈtʃina/).

Headedness, and treatment of SUBSTEMS

As outlined in Section 4.4.3 of Chapter 4, most SUBSTEMS (shown here in square brackets) are added to the existing STEM (shown here in round brackets) by agglutination, e.g. in monte[luk](ast) with 'parent' STEM -ast and 'child' SUBSTEM -luk-, as in Figure 6.5. By reading right to left, the user can first parse the name as an antiasthmatic (with the '-ast' STEM), then further sub-categorise as a leukotriene receptor antagonist (with the '-luk-' SUBSTEM). By agglutinating STEMS in this way, names show a clear hyponymic relationship between 'parent' and child' concepts and users can work right to left to decipher meaning. These names are right-headed, because they show their semantic head at the end of the name, and they make up around 69.0% of the subset of names analysed for headedness (312/452 names). This process is replicated in Italian because names are transliterated so closely: monte[luk](ast).

[26] 774 names in the dataset analysed as part of Chapter 4 are listed in the WHO Stem Book and contain a STEM compliant with WHO naming guidelines. 761 of these names have an Italian counterpart in the trilingual corpus.

Figure 6.5 Hyperonymic STEM building in montelukast (reproduced from Figure 2.3)



Around a third of names in English are not right-headed (~30%). In these cases, the position of the STEM in English is the same in Italian, so the Italian names are also not right-headed. A handful put their STEM at the beginning of the name ('left-headed'), such as cefixime > cefixima, with cef > cef. Others are FREEFIX STEMS, which can appear anywhere in the name without changing its meaning or the STEM taxon to which it belongs. For example, ergocalciferol > ergocalciferol, with calci > calci. Their SUBSTEMS attach in various ways, even as a wrapper on either side of the STEM, as in ator[va](stat)[in] > atorvastatina, with parent STEM stat and SUBSTEM -vastatin. Because these names are not right-headed, they still display their semantic head (endocentric), but they are difficult to parse because their taxonomy is not displayed in a predictable order.

A small number of SUBSTEMS are not agglutinated to their STEM, but rather replace it (17 names)[27]. For example, lami[vudine] > lamivudina (replacing parent STEM -uridine / uridina), and sot[alol] > sotalolo (replacing parent STEM -olol). This replacement may involve a change in position (e.g. [ni]car[dipine] > [nicardipina] (replacing parent STEM nitro-). It rarely involves the introduction of a negative SUBSTEM (e.g. tegafur > tegafur, replacing parent STEM -uridine / uridina). Consequently, these compounds no longer contain a head but rather a hyponym of the head in the form of the SUBSTEM, causing them to adopt an exocentric neoclassical compound structure. In all of these cases, the order of STEMS and SUBSTEMS is maintained in Italian translation, indicating that names are treated as full units and not as multi-componential terms.

[27] (delete?) Replacement: 3 uridine; 1 pred-olone; 2 orphan; 2 olol; 3 drine; 4 azepam; 2 ster

In all cases, FREEFIX STEMS (e.g. calci, gab, gli, prost, vir) do not change position in translation. For example, the FREEFIX STEM vir > vir in aciclovir > aciclovir , enfuvirtide > enfuvirtide, maraviroc > maraviroc. There is no evidence of using word formation processes in translation into Italian, beyond naming of enzymes (13 names, e.g. idursulfase > idursulfasi) and sugars (five names, e.g. lactulose > lattulosio), presumably because these terms exist independently of the INN nomenclature.

Motivation of STEMS and the impact of transl(iter)ation

Chapter 4 explores the *motivation* of STEMS, i.e. how STEMS are formed. STEMS in English are typically motivated by one of four mechanisms, though these are asserted based on first-hand observation of etymological links, rather than citing any published guidance or literature. The mechanisms are:

1. Abbreviating the name of the pharmacological group;
2. Forming a partial acronym of the name of the pharmacological group;
3. Abbreviating the chemical name; and
4. Abbreviating an existing INN.

In translation out of English these links may be broken and this will lead to a loss in semantic motivation of the STEM. Each mechanism will be discussed in turn, drawing examples from Table A8.8 of frequent STEMS in English and Italian.

Type 1: Abbreviating the name of the pharmacological group

The first mechanism is by (ostensibly) abbreviating part of the name for the pharmacological group[28]. For example, -dil for vasodilators (minoxidil > minoxidil), stat for enzyme inhibitors (also known as statins, e.g. fluvastatin > fluvastatina); prost for prostaglandins (e.g. tafluprost > tafluprost); -stim for colony stimulating factors. In translation, some of these STEMS may be decoupled from their semantic motivation in English, because the name for the pharmaceutical group will be different in Italian. It depends on the extent to which Italian uses neoclassical elements in pharmaceutical terms, because pharmaceutical language heavily uses neoclassical elements, e.g. vasodilators, with the combining form *vaso-* from Latin *vas* , vessel'.

I explored the translation of some of the pharmacological groups that are used to form the STEM, and found that many of them share etymological roots with the English term. Many of

[28] 107/452 names in frequent STEM taxa

the Italian terms describing the pharmacological group contain the translated STEM, so they replicate the semantic motivation when translated, with examples in Table 6.9. In any language and context it is difficult to gauge the extent to which users of medication names understand technical terms and neoclassical forms, because the expertise of individual users is so wide, ranging (from patients to consultants) and complex. At a minimum at least the semantic motivation of these STEMS is maintained in translation to some extent and they are not completely arbitrary forms to Italian users.

Table 6.9 Examples of STEMS in English and Italian that are formed by abbreviating the name of the pharmacological group

English STEM	***Pharmacological group (English)***	***Italian STEM***	***Italian translation of pharmacological group***
dil	vasodilators	dil	vasodilatori
gab	gabamimetic agents	gab	agenti gabamimetici
gest	progestogens	gest	steroidi, progestinici
gli	(anti)hyperglycaemics	gli	antiiperglicemici
prost	prostaglandins	prost	prostaglandine
sal	salicylic acid derivatives	sal	derivati dell'acido salicilico
stat	statins, enzyme inhibitors	stat	statine, inibitori dell'enzima
stim	colony-stimulating factors	stim	fattori stimolanti le colonie
vir	antivirals	vir	antivirali

Type 2: Forming a partial acronym of the name of the pharmacological group
The second mechanism is to use the name of the pharmacological group to form an acronym[29]. These may use the first parts of each word or a combining form inside the word, such as -mab for Monoclonal AntiBodies (e.g. infliximab > infliximab; tomalizumab > omalizumab). They also combine abbreviations and single letters, such as -imus for IMmUnoSuppressants (tacrolimus > tacrolimus) and -rafenib for RAF (Rapidly Accelerated Fibrosarcoma) kinasE iNhIBitors; -tinib for Tyrosine kinase INhIBitors.

I explored the translation of these STEMS in Russian: -imus; -mab; -rafenib; and -tinib. Some of the partial acronyms in English can be formed in the same way in Italian. For example, *immunosuppressants* is translated into Italian as *immunosoppressori* and so the partial

[29] 102/452 of names in frequent STEM taxa

acronym STEM -imus can still be derived from the Italian pharmacological group. *Monoclonal antibodies* in Italian is *anticorpi monoclonali*, which cannot be used to derive the partial acronym -mab, so the Italian STEM -mab will not hold the semantic link and instead would form the acronym *acm. The STEMS -tinib and -rafenib will also not maintain their semantic link because Italian word order is different from that of English. In both cases, the phrase is reversed: e.g. *tyrosine kinase inhibitors* > *inibitori della chinasi della tirosina*, lit. 'inhibitors of the kinase of tyrosine'.

Type 3: Abbreviating the chemical name

The third STEM formation mechanism is to abbreviate the chemical name, e.g. -cillin for 6-aminopenicillanic acid derivatives, in benzylpenicillin > benzilpenicillina[30]. As in the first mechanism of abbreviating the pharmacological group, the semantic motivation is likely maintained in translation as long as Italian uses a neoclassical form for the chemical compound or active ingredient. For the frequent STEMS in Table A8.8, most of the chemical names use the same neoclassical combining form in translation, so the semantic motivation of the STEM is maintained. In the case of -mycin, this is derived from the genus name in Latin, *Streptomyces*, also used in Italian, so the semantic link is maintained. Examples are presented in Table 6.10.

Table 6.10 Examples of STEMS in English and Italian that are formed by abbreviating the chemical name

English STEM	***Pharmacological group (English)***	***Chemical from which the STEM is derived***	***Italian STEM***	***Italian translation of chemical name, with pronunciation***
calci	vitamin D analogues/derivatives	calcium	calci	calcio
-cillin	antibiotics, 6-aminopenicillanic acid derivatives	6-aminopenicillanic acid	cillina	acido 6-aminopenicillanico
cef-	antibiotics, cefalosporanic acid derivatives	cefalosporanic acid	cef-	acido cefalosporanico
-mycin	antibiotics, produced by Streptomyces strains (see also -kacin)	Streptomyces	-micina	Streptomyces
-retin	retinol derivatives	retinol	-retin	retinolo

[30] 55/452 of names in frequent STEM taxa

-tide	peptides and glycopeptides (for special groups of peptides see -actide, -pressin, -relin, -tocin)	peptides, glycopeptides	-tide	peptidi, glicopeptidi

Broadly speaking, neoclassical combining forms are used heavily in the Italian language since they have a shared ancestor in Latin and strong historical links with Ancient Greek. Several neoclassical combining forms are used in INNs, such as *chlor*, *hydro*, and *meth*. These are consistently translated into the Italian neoclassical forms *clor* (e.g. chloramphenicol > cloramfenicolo), *idro* (e.g. hydrocortisone > idrocortisone), and *met* (e.g. methyldopa > metildopa). The combining form oxi, as in *ossigeno* ('oxygen') is translated variably as either *oxi* or *ossi* (e.g. digoxin > digossina; doxycycline > doxiciclina).

One STEM is a standard chemical suffix: -ase, for enzymes. The SUFFIX STEM for enzymes in Italian is *-asi*, e.g. amylase > amilasi; lactase > lattasi. In these names, the STEM is derived from the accepted chemical suffix. For example, idursulfase > idursulfase and urokinase > urochinasi.

Type 4: Abbreviating an existing INN

Many INNs are ostensibly named after the first INN in that STEM family (because the INN is mentioned in the definition of the pharmacological group)[31]. It is unclear in these cases what the motivation for that first INN was, and it may be that it is arbitrary. For example, -azepam is the STEM for diazepam derivatives (diazepam was approved in 1965) and -oxetine for fluoxetine derivatives (fluoxetine was approved in 1976). In these cases, the first INN in the taxon serves a semantic function and is the relative motivation for others in the taxon. This motivation will be maintained in translation because Italian names are close orthographically to English INNs, and the majority of STEMS are either the same or differ by a single letter at the end. For example: miconazole > miconazolo; sumatriptan > sumatriptan. In some cases the first INN has since been discontinued, e.g. ibufenac in the -ac taxon, for "anti-inflammatory agents, ibufenac derivatives" e.g. etodolac > etodolac; and piperazine in the -izine taxon for "diphenylmethyl piperazine derivatives", e.g. hydroxyzine > idroxizina.

Unclear semantic motivation

[31] 83/452 of names in frequent STEM taxa

In many cases the semantic motivation of the STEM is unclear, as in -pril for "ACE inhibitors", e.g.> ramipril > ramipril; and -terol for "bronchodilators, phenethylamine derivatives", e.g. formoterol > formoterolo[32]. The STEM is in the same position and form in translation, but semantic motivation is just as unclear.

6.4.2.3 Relational features of Italian generic names

Chapter 4 explores ten linguistic features of INNs that are relational, i.e. they measure how names interact in the dataset. Because STEM information is not available for Italian names, five of the ten measures are possible to analyse in Italian: the frequency of two-letter and three-letter strings at the beginning and endings of names (so-called *ngrams*) and neighbourhood density. Neighbourhood density is the number of names within a fixed orthographic distance of a target name, in this case with a minimum similarity score of 33%. The trigram-2b score is used, which counts the proportion of shared three-letter strings in two names, adding two spaces to the beginning of each name to weight the beginning, as in Lambert et al. (1999). For example the neighbourhood of fluconazole includes: flucytosine, ketoconazole, miconazole and eight other names. Names sharing their first letter have been linked with LASA error (in Shah et al., 2017), as have names with a high neighbourhood density (Lambert et al., 2001b).

<u>Beginnings and endings of Italian INNs</u>

As in the analysis of English names in Chapter 4, the frequency of two-letter strings (bigrams) and three-letter strings (trigrams) is explored at the beginnings of Italian names and results compared between those for English names. The most frequent (top ten) initial and terminal trigrams in Italian and English INNs are given in Table 6.11. Appendix sections A23 to A26 give supplementary information on distribution and frequency of initial and terminal ngrams.

The top ten two-letter strings at the beginning of names make up a quarter of Italian names, e.g. *me* in metformina and *cl* in clopidogrel. The most frequent initial bigram of INNs in Italian is *me* and the most frequent trigram is *clo*. The distribution of initial strings is comparable between Italian and English, indicating very little difference in the way names are started in both languages (Italian names have 547 unique initial trigrams, and English names have 567). Names sharing the first letter and sharing strings of letters have been linked with LASA

[32] 92/452 of names in frequent STEM taxa

error (Shah et al., 2017) so if the distribution of initial strings is similar in Italian to that of English then the risk is relatively similar in Italian generic names.

The most frequent two-letter ending of INNs in Italian is *na* (e.g. argipressin > argipressina) and most frequent three-letter ending is *ina* (e.g, thrombin > trombina), appearing in more than 30% of Italian names. This is similar to English names, in which *ine* appears in more than 20% (though is phonemically distinct). As outlined in Chapter 4, there is more diversity in the beginnings of English names than in their endings. English names' endings are more homogeneous than their endings. In English names there are 567 distinct beginnings (initial trigrams), compared with 132 distinct endings (terminal trigrams). Over a quarter of English names end in *ine* or *ide*. This homogeneity in endings is stronger in Italian, because so many of the English names ending in *in* and *ine* become *ina*. There are 114 distinct terminal trigrams in Italian.

Table 6.11 Initial and terminal trigrams in Italian and English (top ten most frequent, in descending order of frequency)

	Initial trigrams		***Terminal trigrams***	
Italian	clo- flu- pro- met- car- tri-	cef- fen- lev- ben- idr- tra-	-ina -olo -ide -one -mab -nib	-vir -ene -tan -ato -ril
English	flu- pro- car- met- chl-	clo- tri- cef- lev- hyd-	-ine -ide -one -mab -nib	-cin -vir -ole -tin -tan

Neighbourhood density of Italian generic names

Neighbourhood density in Italian is similar to that of English. Results are compared between Italian and English in Figure 6.6, and distribution is given in the Appendix in Table A8.7. The majority of names have between one and five neighbours (559/1024, 54.59%), as in English (607/1046, 58.0%). Many more Italian names than English names have more than ten neighbours (119 compared with 80). For example, doxylamine > doxilamina with 11 neighbours, including difenidramina, doxapram, ketamina, tiamina. This is likely due to the smaller orthography used in Italian, excluding unnecessary letters such as <y>. Italian

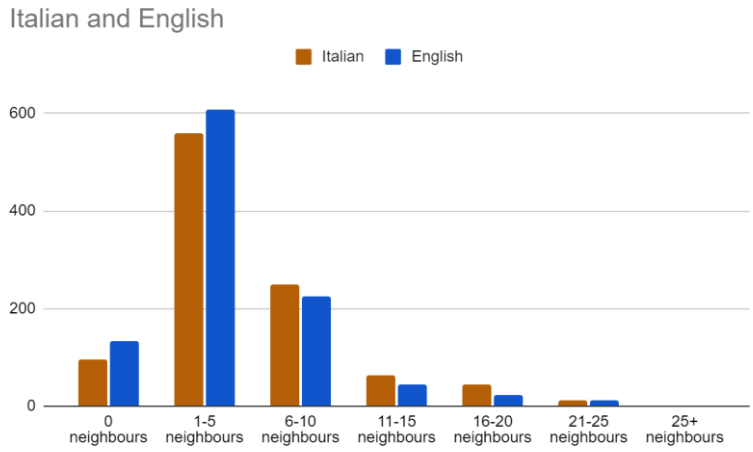
names for highly frequently prescribed medications (as in Table 6.7) all have at least one neighbour, but up to 15 neighbours (clozapina, with: asenapina, clorexidina, clobazam, clomifene, loxapina, olanzapina).

Italian neighbourhood density is closely correlated with that of English (Pearson rho=0.801, p<0.001). Many names have an identical similarity neighbourhood in Italian as in English, e.g. ramipril < ramipril, both with the same four orthographic neighbours, including moexipril < moexipril and imipramina < imipramine.

Some differences in neighbourhood are due to a difference in word length, which skews the calculation used to calculate orthographic similarity (the Dice coefficient, which is the proportion of shared strings to total strings). For example, ibuprofen and bupropion are both shorter than ibuprofene and bupropione. Since the cut-off is 33%, in Italian these are not neighbours (30.0% similar) whereas in English they are (33.3% similar). In other cases, the addition of a vowel on the end creates another shared three-letter string, also skewing the calculation. For example, colecalciferol and calcipotriol have a trigram-2b score of 32%, but colecalciferolo and calcipotriolo have a score of 35% because the Italian names share an ending as well as the STEM calci.

At times, the phonemically-driven orthography used in Italian compared with English means that Italian names share more strings so have more neighbours. For example, fentanil and fenilefrina are 40% similar, but fentanyl and phenylephrine are 0% similar, because Italian uses <f> instead of <ph> and <i> instead of <y>. In this case, a smaller orthography and avoiding prohibited letters increases similarity in the nomenclature.

Figure 6.6 Distribution of neighbourhood density in Italian



6.5 Discussion and contrastive overview

This chapter has presented case studies of nomenclatures in two languages. The first, International Nonproprietary Names in Russian, which are designated by the WHO; the second, Italian Common Names that are designated by the Official Pharmacopoeia of Italy. In this discussion section, findings are summarised from the case studies, and contrast findings between Italian, Russian, and English. The implications of findings from the statistical analysis in the previous Chapter 5 for these nomenclatures are also considered. This is answering Research Question 3: *How can Russian-language and Italian-language pharmaceutical nomenclatures be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?*

I analysed 1,046 Russian-language INNs in the case study in Section 8.2, and 1,024 Italian-language generic names in the case study in Section 8.3. Russian names are a little **shorter** than English names and use a smaller orthography than in the general language, excluding five letters from the Russian alphabet. In contrast, Italian names are little **longer** than English names and use all the letters in the Italian alphabet, plus additional characters used in loanwords. In Russian, transliteration closely follows the ISO 9:1995 standard and in Italian almost half of the names are the same as their English counterparts. Prohibited letters are treated using a similar approach in both languages: render phonemically where an equivalent sound exists in the target language, and mimic how the neoclassical form is treated in other words in the target language. So when a morph exists as a combining form in other words in the target language, transliteration/translation is consistent with that form.

English names that end in a silent <e> are treated variably by either not transliterating it in Russian, or by substituting it in Italian with another vowel such as <a> or <o>. English names that end in a consonant maintain this ending in Russian, but in most cases gain a terminal vowel in Italian. Most Russian names therefore will be treated as masculine and none as feminine, and around a third of Italian names will be treated as feminine but an eighth as masculine (the remainder end in an <e> or a consonant so could be treated as either). Dropping the terminal <e> initially suggests that transliteration into Cyrillic is driven by phonemics of the English names. However, in translation to Italian most names gain a syllable, suggesting an overall strategy of aligning names with the phonotactics of the target language, because many Russian words end in a consonant, while most Italian words end in a vowel.

Since Italian orthography has a more regular relation to contemporary pronunciation, terminal <e> in Italian names forms a syllable, so even English names that do not undergo an orthographic change will still undergo a phonemic change and potentially a stress change too. The reduction of names into Russian by removing the terminal <e> does not reduce the syllabic value of Russian names.

I drew up a list of Russian and Italian STEMS based on observation of how the STEM in English has been transl(iter)ated into Russian, or translated into Italian. As in the rest of the name, the transliteration of the STEM into Cyrillic closely follows the ISO 9:1995 standard. Bar the loss of the final silent [e], there is little change to STEMS in translation to Russian. Because of the vowel that is often added in translation into Italian and because most STEMS are SUFFIXAL, many STEMS change form both orthographically and phonemically in translation to Italian. STEMS do not change position in translation to either language, even when the STEM is a FREEFIX. There is no clear evidence that INNs are treated as decomposable morphological units at all, beyond the fact that they have internally-consistent forms within STEM taxon. This consistency may be explained by there being a strategy of maintaining the form of the STEM, or because the names are treated on a whole-word basis, as SUBSTEMS agglutinate to, or replace, their hyperonymic STEM in the same way in Russian as they do in English. It could also be explained by a strict adherence to a transliteration protocol. The exception is the chemical suffixes for enzymes (-ase) and sugars (-ose), which are translated in the conventional way as other forms in the target language with those suffixes and not phonemically transliterated into Russian or calqued into Italian, and neoclassical combining forms such as chlor and meth, which are internally consistent.

English names with STEMS that are formed by abbreviating the pharmacological group generally maintain that semantic motivation in translation into both Russian and Italian, as do those named after a chemical group or substance, and those named after an existing INN. STEMS formed by taking parts of the pharmacological group to form a partial acronym are likely to lose semantic motivation, because of differences in word order and use of native combining forms in the target languages instead of neoclassical combining forms (see Georgieva, 2013; Puato, 2011). In respect to the semantic motivation of STEMS, the impact of translation is similar between Russian and Italian.

Due to close transliteration into Cyrillic, the most common ngrams at the beginnings of names are the same in Russian as in English, and similar in Italian as in English. Due to dropping the final <e>, the diversity of Russian three-letter endings is somewhat higher than in English but not in two-letter endings, and pronunciation does not change. The distribution of initial and terminal trigrams is illustrated in Figures 6.7 and 6.8 below.

Figure 6.7 Distribution of initial trigrams in Russian and Italian

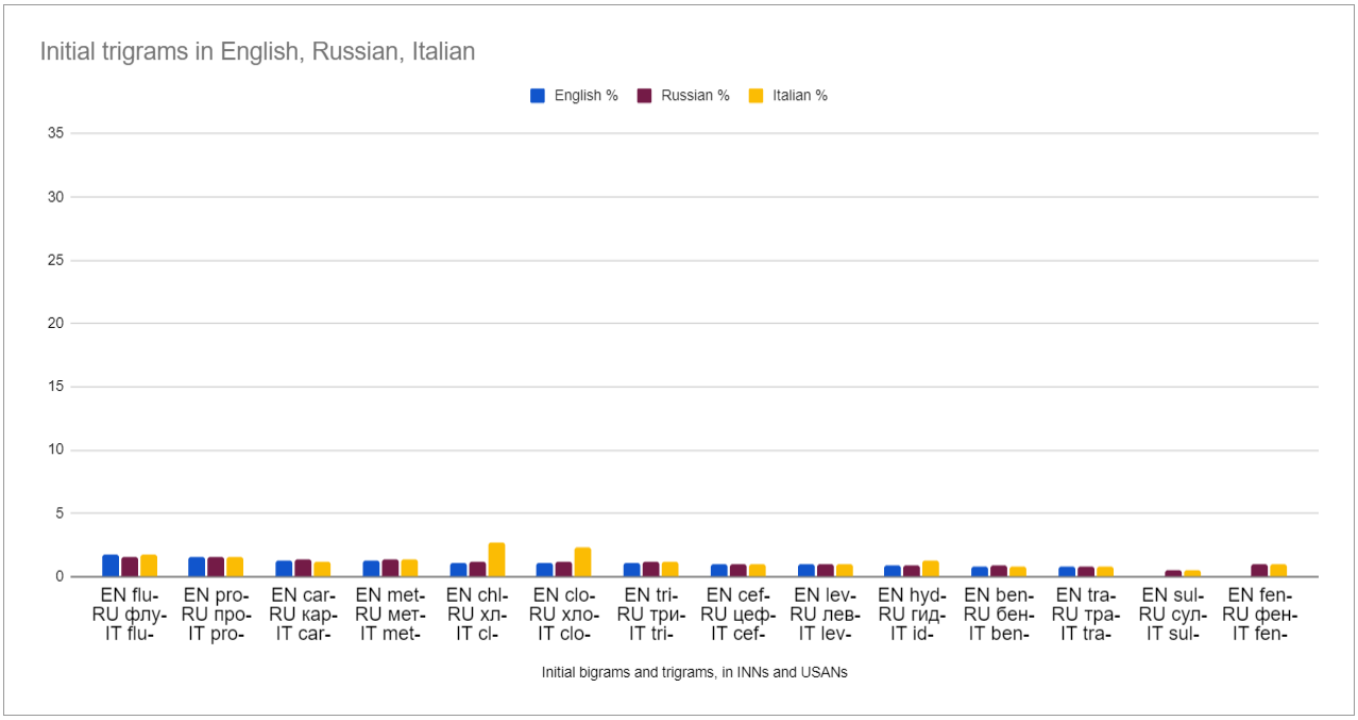
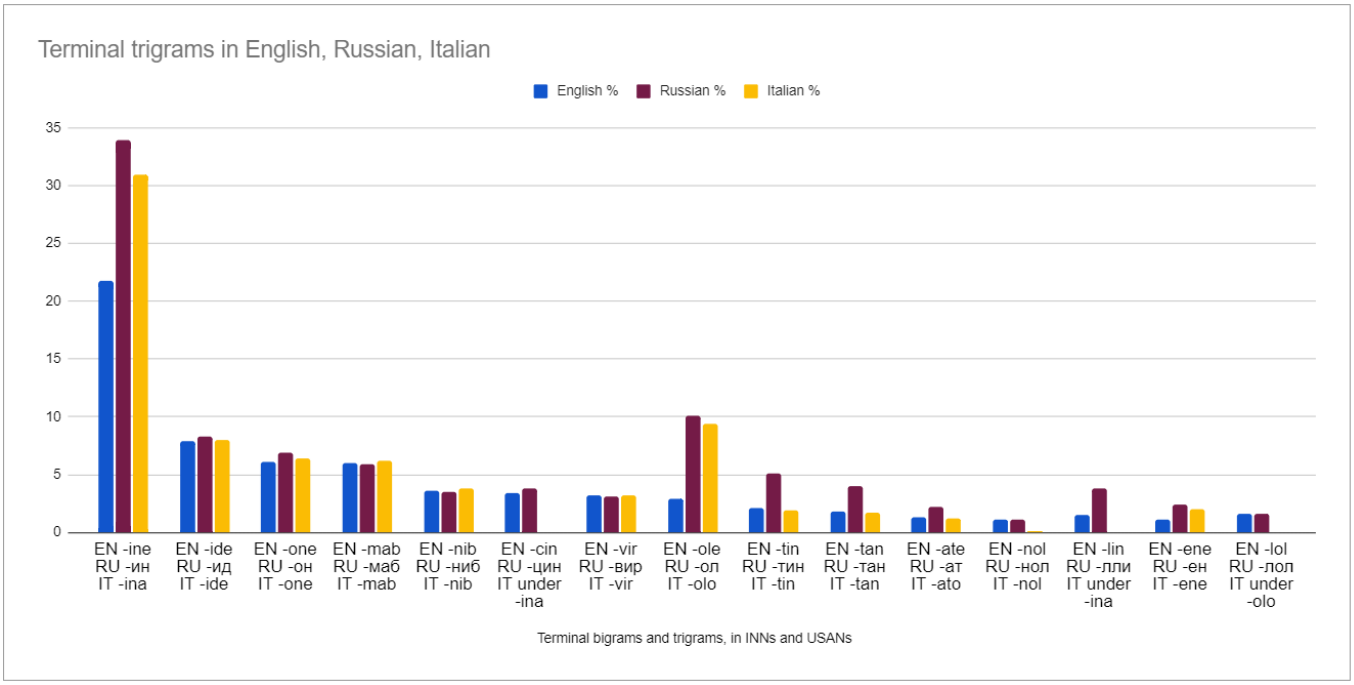


Figure 6.8 Distribution of terminal trigrams in Russian and Italian



Because the transliteration strategy is consistent in Russian (compliant with the ISO standard) and the translation strategy is consistent in Italian, relative similarity between names in each dataset is stable. This means that neighbourhood density in Russian and Italian names is similar to that of English names, with the majority having between one and five neighbours. More names in both target languages have at least one neighbour, possibly due to the phonemic transliteration of prohibited letters such as <ph> and <y>.

6.5.1 Global connotations of drug names

The translation of drug names into numerous languages from diverse language families can often result in unintended connotations in these target languages that may not be present in the source language. While it is not impossible to entirely avoid connotations in all languages, efforts are made by the WHO INN Expert Committee to minimise the risk of particularly offensive names (SoINN, 2020b; WHO, 2017a).

The case studies in this chapter include lists of pharmacological STEMS and their *assumed* equivalents in Russian and Italian (presented in full in Table A8.1). There are a handful of STEMS that may carry alternative meanings in the target languages. For example, the Russian STEM прост connotes *простой*, 'simple'.

Through personal communication with an advisor to the WHO INN Committee, Dr. Jeffrey Aronson from the University of Oxford, I learned of a case where a name was approved as an INN but was later changed due to multiple complaints from Swedish speakers. To Swedish speakers, the drug name upifitimab sounds like a colloquial phrase in Swedish, *opp i fitta mi,* 'up in my pussy'. This proved particularly offensive given the drug's therapeutic indication for ovarian cancer. While this case stands out, there are likely many more examples of suboptimal or unintended connotations in target languages. Dr. Aronson also highlighted that Japanese often presents a challenge in avoiding unintended meanings, and more connotations are likely overlooked due to a lack of representation of other languages on the Committee. These instances underscore the importance of carefully considering the linguistic implications of drug names across different languages to avoid unintended connotations and potential offensive undertones.

6.6 Implications of findings from RQ2 in Russian and Italian

This section answers the second part of RQ3, namely: *"... what are the implications of findings from RQ2 for these languages?"* Implications of the findings from the statistical analysis in Chapter 5 for the generic drug nomenclatures in Russian and Italian are considered.

In the statistical analysis in Chapter 5, thirteen linguistic variables were explored for their association with the binary outcome of whether the name has been reported in a LASA error, and two non-linguistic confounders (prescribing frequency and year of approval). Seven of these remain in the final model, but only one has sufficient empirical evidence to warrant any suggestions for policy changes. The other six are tenuously associated with error based on inferences of the results, but need further research to understand this association. It was not possible to collect equivalent data in Russian and Italian for all of these, so they will be discussed in turn as the data availability allows. Each of the associated variables are contextualised in the Russian and Italian nomenclatures, regardless of their veracity as a predictor of error.

6.6.1 Confounders: prescribing frequency and year of approval

Low prescribing frequency and recent year of approval both have a protective effect against LASA errors (both $p<0.001$). The two confounders used in the logistic regression models were found to be highly correlated with LASA error. Names that have a high prescribing frequency and were approved a long time ago are more likely to be reported in a LASA error, presumably because there are more opportunities for confusion with the increased usage.

Prescribing frequency may differ slightly between nations due to regulatory, lifestyle, environmental, and cultural differences between the UK, Russia, and Italy (e.g. (Zito et al., 2008) but substantive differences are not anticipated. The year in which names are approved is the same between Russian and English INNs as they are published at the same time by the WHO, and there will be up to a three-year difference between Italian and English, as the European Pharmacopoeia is published every three years. Thus prescribing frequency and year of approval are likely to influence the risk of LASA error in Russian and Italian-language names.

6.6.2 Orthographic depth, or presence of prohibited letters

The presence of prohibited letters in names was found to have a protective effect against LASA error, although the direction of the effect is not certain and the association is tenuous and likely confounded by year of approval (p=0.067, OR 0.629 [0.380, 1.027]). This was adopted as a proxy of orthographic depth - the guideline to avoid these letters is designed to facilitate translation and international understanding of the name by promoting a one-to-one correspondence between graphemes and phonemes. The guidelines are only published in English, for the English nomenclature, so they do not apply to the other languages, and it is therefore challenging to discuss the implications of this finding for Russian and Italian names. In the dataset, Russian names draw upon a smaller set of letters than in the general language (26 letters versus 33) and Italian names extend the orthography of the general language (24 letters versus 21) by adding <k>, <w>, and <x>. From this it could be surmised that orthographic depth is deeper in Russian and more shallow in Italian, than in English names, thus the protective effect is amplified in Russian and weakened in Italian. However, the presence of prohibited letters is not a perfect proxy measure for orthographic depth and the effect size is unclear for this factor.

6.6.3 Characteristics of STEM taxon type (Morph-09)

Under the typology first conceived in Bryan (2015) and developed further in this thesis, names were categorised into one of four taxon types: simple, regular, complex, irregular. When the variable was expanded to include listedness, compliance, and STEM taxon size, and treated as an unordered factor, some evidence was found of a difference in LASA error between regular and irregular taxon types (Morph-09, p<0.001 all factors). Although STEM information is not available for the other languages under study here, the position of STEMS in translation and transliteration does not change for any of the names in the dataset so the type of STEM taxon will also be unchanged in the Russian and Italian nomenclatures. Because STEMS do not change their position in translation into Russian or Italian, and are a type of morphological calque, this effect should be the same for these nomenclatures.

However, the taxon typology developed in this thesis is quite Anglocentric, conflating right-headedness with 'regularity'. Names that are exocentric (do not have a semantic head) or are left-headed (showing their semantic head at the beginning of the name) are deemed irregular, in type 4. Nominal compounds are overwhelmingly right-headed in Germanic languages but less so in Romance and Slavic languages (Guevara & Scalise, 2009).

In Russian, compounds are most commonly right-headed, e.g. *подоконник*, 'windowsill', lit. 'under ledge', including in neoclassical compounds, e.g. *микроволновой*, 'microwave' (adj.). In Italian, left-headedness is dominant, although right-headedness is typical in neoclassical compounds (e.g. *cardiologia*, 'cardiology', lit. 'the study of the heart') with many examples in the general language (e.g. *telecomando*, 'remote control', a type of *comando*, 'control') (Cacchiani 2020; Marelli and Luzzatti 2012). In contrast to Germanic and Slavic languages, Romance languages have higher rates of exocentric compounds and left-headedness, though this tends to be used more for lexicalisations of syntactic constructions (e.g. *capostazione*, 'station master', a shortening of *capo della stazione*) or items of medieval origin (e.g. *pomodoro*, 'tomato', lit. 'apple of gold') (Dardano, 1978; Guevara & Scalise, 2009; Marelli & Luzzatti, 2012). Since right-headedness is not as dominant in Russian as in English and not dominant at all in Italian, it may be that the protective effect of irregularity is dampened in these languages. To put it another way, features such as left-headedness and exocentricity are unusual in English and it appears that this difference makes Russian and Italian names less liable to confusion - if these features are not so unusual in Russian and Italian then the benefit will be lost. However, the names are partial neoclassical compounds, typically right-headed in Russian and Italian as well as in English, so this may mean that the differential effect is maintained in translation.

6.6.4 Initial and terminal trigram frequency (Rel-02 and Rel-04)

Diversity in the beginning and end strings has a protective effect against LASA error (initial trigrams $p<0.001$ and terminal trigrams $p=0.017$), because this reduces the number of potential pairwise shared strings within a nomenclature. Russian and Italian scores were substituted for the English scores of these variables, with negligible difference to performance (forward selection model, English AIC 845.803; Russian AIC 841.934; Italian AIC 841.119) and the scores are significant (e.g. Russian initial trigrams, $p<0.001$, OR 1.084 [1.039, 1.130]). The distribution of initial trigrams in Russian and Italian is very similar to that of English names, so the effect should be on par. Russian terminal trigrams are more diverse than those in English, which might amplify the effect and reduce LASA errors. On the

contrary, Italian terminal trigrams are *less* diverse than those in English, which might dampen the effect and increase error rate.

6.6.5 Alignment with the ATC classification system

There was a very unreliable association between the ATC alignment of STEM taxa in the nomenclature and LASA error (p=0.096, OR 1.712 [0.898, 3.205]). ATC codes are international, so are the same for names in other languages. Since guidance on STEMS was only available in English, STEMS in Russian and Italian were inferred from the names themselves and are given in Sections 8.2 and 8.3 respectively. Membership of a STEM taxon is not changed for any name in the dataset so it can be inferred that the ATC alignment scores are the same in Russian and Italian as they are in English.

6.6.6 Neighbourhood frequency

Having a neighbour that is prescribed more frequently than itself was found to be a possible predictor of LASA error, although the findings are not certain (p=0.081, OR 1.367 [0.963, 1.943]). It is not possible to explore the implications of this finding for nomenclatures in Russian and Italian because prescribing data was not collected for their respective countries in this study.

6.6.7 Semantic transparency (Rel-10)

There is some evidence to suggest that increased semantic transparency, measured by positional family size, has a protective effect against LASA error (p=0.030, OR 0.496 [0.261, 0.930]). As with taxon type, the STEMS are in the same position in translation and were closely transliterated and translated, so the measure should be unchanged in the Russian and Italian nomenclatures. Semantic transparency is an internal measure of consistency in the way that names in a STEM taxon communicate membership of that taxon, and is not subject to top-down judgements of what is normal or typical, e.g. right-headedness. If all the names in one STEM taxon are left-headed, and all the names in another STEM taxon are right-headed, all names are judged to be semantically transparent because they are relationally predictable. Because STEMS are observed in Russian and Italian in this thesis and are not official STEMS (published by the respective naming bodies), it is not possible to be certain of STEM taxa - however, in no name in the dataset does a STEM change position in translation to either Russian or Italian (see earlier discussion of *morphological calques*), so we can surmise that semantic transparency scores are the same in these languages. However, as discussed above, it may be that because right-headedness is not as dominant in Russian

and Italian as it is in English, right-headed compounds have a 'foreignising' effect, which may serve to reduce semantic transparency (Venuti, 2017).

Chapter 7: Discussion and conclusions

This thesis addresses the problem of Look-Alike Sound-Alike (LASA) errors between medication names. LASA errors are a significant concern in patient safety with potentially serious consequences if gone unnoticed. The literature on LASA errors generally falls into three categories, examining: the likelihood of errors, the opportunities for errors, and the seriousness of errors. The focus of this thesis is on linguistic factors that increase the likelihood of LASA errors, with an emphasis on solutions during the stage prior to the name being approved. Under James Reason's 'system approach', which views errors as inevitable and humans as fallible, error is driven by latent conditions in the system that need to be identified and mitigated. This thesis aims to pinpoint *linguistic latent conditions* contributing to the risk of LASA errors. Addressing critical gaps in the literature, the study focuses on the often-overlooked semantic features of nomenclature, linguistic properties of generic names, and the understudied areas of LASA errors and pharmaceutical nomenclatures in languages other than English. Overall, this thesis enhances understanding of LASA errors by exploring linguistic features of generic medication names and contextualises findings within an international and multilingual context.

7.1 Outline of research aim and questions

This thesis has worked to identify linguistic latent conditions in international medication names, specifically those that contribute to the risk of LASA errors. These linguistic features are considered within a global and multilingual context. There are three analytical components to the thesis, making up Chapters 4, 5, and 6.

In Chapter 4, linguistic features of International Nonproprietary Names (INNs) in English are characterised by the creation of a taxonomic database, answering Research Question 1: *How can the INN nomenclature be characterised in linguistic terms?* A basket of measures of orthography, morphosemantics, and relational aspects of the nomenclature are explored. In Chapter 5, the association between linguistic features analysed in Chapter 4 and the likelihood of a name being involved in a LASA error is measured, answering Research Question 2: *Are any of the linguistic features described in answer to RQ1 associated with the occurrence of LASA error?* The association of linguistic features with error is modelled, in order to identify linguistic latent conditions. In Chapter 6, a case study approach is adopted

to characterise nomenclatures in two languages: Russian-language INNs and Italian Common Names. Furthermore, implications of findings from Chapter 5 are considered, looking at how linguistic latent conditions manifest in translation and/or transliteration into Russian and Italian. Chapter 6 answers Research Question 3: *How can Russian-language and Italian-language pharmaceutical nomenclatures be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?* The three analysis chapters are interconnected and build upon each other in a stepwise fashion. First, names are described linguistically (RQ1). Second, statistical analysis determines whether these linguistic features are associated with LASA error (RQ2). Third, findings are considered through a global and multilingual lens (RQ3).

The thesis focusses on both potential LASA errors (medication name pairs with similarities) and examples of pairs of medication names that have been confused and are deemed 'LASA pairs' (ISMP 2019). It specifically examines generic medication names: INNs designated by the WHO in English and Russian, and names designated in Italian by the Farmacopea Ufficiale Italiana (FUI, Ii. *Official Pharmacopoeia of Italy*). The study looks at LASA errors between similar names, rather than issues related to packaging or labelling alone. Orthographic and phonological similarities are taken into account, while errors related to individual letter glyphs (e.g. 2iu/21u) are excluded. The dataset is restricted to names for medications that are authorised for use in the UK, as it considers the likelihood and risk of LASA error – unauthorised medications have an extremely low risk of being reported in a LASA error. Research Questions 1 and 2 are explored for the English nomenclature only, as WHO naming guidance is provided solely in English, and available lists and case reports of LASA errors are primarily in English, complemented by case studies of Russian and Italian names in Chapter 6, answering Research Question 3.

7.2 Research questions 1 and 2: Findings

Findings for Research Questions 1 and 2 are discussed in this section. This section synthesises findings on linguistic features characterised in Chapter 4 in creating the taxonomic database, broken down by orthographic, morphosemantic, and relational features. This is accompanied by a discussion of those features found to be associated with a difference in LASA error risk, highlighting where appropriate avenues for future research and suggestions for manufacturers proposing INNs and the WHO INN expert committee. Discussion is complemented by summary tables for each of the three types of features (Tables 7.1, 7.2, and 7.3). Findings are accompanied by a discussion of how these findings confirm or contrast with findings in the literature, and how the findings contribute to the knowledge base of the field.

There were four *a posteriori* assumptions of linguistic features that would be found to be associated with LASA error, based on findings in the literature: names containing high-frequency initial strings[33] and high-frequency terminal strings[34] (either bigrams or trigrams, based on Shah et al. (2017); names with a high neighbourhood density[35] and with a more frequently-prescribed neighbour[36] (based on Lambert, Chang, & Gupta, 2003; Lambert, Chang, & Lin, 2003). Using a backward elimination stepwise model of binary logistic regression, followed by stability investigations, seven linguistic variables were identified to be significantly associated with error (the newly merged STEM taxon[37], initial trigram frequency[38], terminal trigram frequency[39], semantic transparency[40], prohibited letters[41], more frequently-prescribed neighbour[42], ATC alignment[43]). Of these seven, most have an unreliable connection to LASA error that should be interpreted with caution but may warrant further investigation, apart from Rel-$02_{\text{initial trigrams}}$. The findings tentatively support three of the four assumptions, but do not find an association between neighbourhood density and LASA error. The significantly-associated variables are potential linguistic latent conditions: factors that are present in the system that may prime the risk of LASA error.

7.2.1 Dominance of confounders: prescribing frequency and year of approval

In the statistical analysis the two non-linguistic confounders dominated the LASA error risk level: prescribing frequency and year of approval. Alone these two confounders explain 20% of the variance in error. Predictably, the risk of a name being involved in a LASA error increases with the prescribing frequency and age of the name (i.e. when it was approved), so it would benefit pharmacovigilance agencies to factor in prescribing frequency and age of the name to highlight at-risk names.

---

[33] Rel-02
[34] Rel-04
[35] Rel-08
[36] Rel-09
[37] Morph-09
[38] Rel-02
[39] Rel-04
[40] Rel-10
[41] Orth-03
[42] Rel-09
[43] Rel-07

7.2.2 Orthographic features of INNs and their association with LASA error

Orthographic features of INNs include measures of character and syllable count (Orth-01 and Orth-02), prohibited letters (Orth-03), and Scrabble score (Orth-04). The findings for these are summarised in Table 7.1. The presence of prohibited letters was explored both to test compliance to WHO naming guidelines and as a proxy for orthographic depth. The findings on name length confirm an earlier study by Bryan et al. (2015), which measured the entire INN nomenclature published up to 2014 (n=7,987), whereas the current dataset includes only names for medications that have been authorised in the UK (n=1,046). Agreement with prior research suggests that the findings are robust and can be generalised to the wider nomenclature. Findings on name length also confirm the findings on name length of USANs from a study by Lambert et al. (2001a). Despite the present thesis and Lambert's study being two decades apart, INNs seem relatively similar to USANs, albeit slightly longer. Findings on the presence of prohibited letters adequately confirm those in previous work (Bryan et al., 2015). Although the previous 2015 study found that a lower proportion of names (13% - 1,036/7,987) contained at least one prohibited letter, it pre-dates the 2017 guideline prohibiting duplicate letters so this figure would be higher if measured again. Findings on orthographic features of INNs update two earlier studies (Bryan et al., 2015; Lambert et al., 2001a) and provide an insight into a subset of the nomenclature: names authorised for use in the UK. It further contributes to the evidence base on drug nomenclature with the first Scrabble score of a set of drug names.

A marked difference was found in the range of scores for these three measures for names approved in the early days of the INN programme (the 1950s and 1960s) and names approved from the 1970s onwards – naming styles have evolved over time. The early names are volatile - they contain a lot of letters that are prohibited by the INN programme and are much longer than later names. This is likely due to the fact that at the beginning of the programme the guidelines were still evolving and were more dynamic (and less understood by regulators and those submitting names for approval). When allowing for differences in length, the Scrabble score stays relatively steady over the years, so complexity in this sense is stable over time.

Table 7.1 Summary of findings on orthographic features of INNs

Var	Descriptive findings	Statistical findings	Policy and regulation	Future research

Orth-01 How many characters does the INN contain?	Mean: 10.6 characters, e.g. candesartan. This confirms the findings of other studies.	Not statistically significant		
Orth-02 How many syllables does the INN contain?	Mean: 4.1 syllables. e.g. latanoprost	Not taken forward for statistical analysis		
Orth-03 Does the INN contain prohibited letters or groups of letters?	18% contain at least one prohibited letter, e.g. pseudoephedrine	Findings are tentative p=0.067, OR 0.629 [0.380, 1.027] The association is overinflated and the variable is present in only 71% of the bootstrapped resamples.	Over time most prohibited letters have dropped out of use, but the use of <k> has remained steady. <k> can be useful in translation and pronunciation, so review of this letter's status in the WHO guidelines is recommended.	Future studies could disaggregate the specific prohibited letters that are associated with error, rather than the binary measure used here.
Orth-04 What would the INN score if played in a Scrabble game?	Mean Scrabble score: 18, e.g. duloxetine. This is the first Scrabble score of a set of drug names.	Not statistically significant		

7.2.2.1 Use of letters and groups of letters that are prohibited by the WHO guidelines (Orth-03)

Three orthographic variables were taken forward for the statistical analysis forming Chapter 5 (with the exception of Orth-02, number of syllables), and one was found to be associated with LASA error: Orth-03 (p=0.067, OR 0.629 [0.380, 1.027]). This is the first research study to measure the association between the presence of prohibited letters (as a proxy for orthographic depth) and LASA error, and as such forms a useful contribution to the literature. Several letters are prohibited for use in INNs, including duplicate letters, <h>, <k>, and <th>, to "facilitate translation and international use" (WHO, 2017a).

The presence of prohibited letters in a name was associated with a reduction in the risk of LASA error. These letters are prohibited by the WHO to facilitate translation and international use of INNs. Prohibited letters include <h>, <th>, <y>, and the more recent addition of duplicate letters such as <ll>. The finding that these letters may have a protective effect on

the risk of LASA error is not sufficiently robust: the confidence interval is broad and crosses over 1, and it was in only 71% of the bootstrapped resamples. Its effect is biased upwards and it was not quite significant to <5%. Certainly, this result should be interpreted with caution.

One explanation for the protective effect of prohibited letters is the fact that it can reduce neighbourhood density. Restricting the use of certain letters that do not have a one-to-one correspondence between phonemes and graphemes (sounds and letters) such as <ph>, and homophonic graphemes such as <y> and <i> that are not present in other languages leads to a simplification of the orthography of names – the pool of letters drawn upon. This was observed in the case studies in Chapter 8 (discussed in full in Section 9.4 upcoming), where phonemic treatment of homophones such as <y> and <i>, and <ph> and <f> reduced difference in translation because these homophonic pairs do not exist in the target languages (Russian and Italian, in this case) so were translated consistently. This led to an increased neighbourhood density in both the Russian and Italian datasets, e.g. English-language amoxicillin has four neighbours (ampicillin, piperacillin, benzylpenicillin, pivmecillinam), compared with Russian-language амоксициллин with eight neighbours (феноксиметилпенициллин, бензилпенициллин, флуклоксациллин, ампициллин, пиперациллин, доксициклин, тамоксифен, пентоксифиллин) and Italian-language amoxicillina with nine neighbours (amorolfina, ampicillina, benzilpenicillina, doxiciclina, flucloxacillina, pentoxifillina, fenossimetilpenicillina, piperacillina, pivmecillinam). The variable neighbourhood density in itself was not significantly associated with the risk of LASA error, but in earlier studies neighbourhood frequency was found to have an interactional effect with neighbourhood density in that when neighbourhood density was high, neighbourhood frequency affected recognition and accuracy (Lambert, Chang, & Gupta, 2003). It may also be that the use of digraphs such as <ph> actually add to the *distinctiveness* of names and trigger recognition and recall.

However, year of approval is likely to be confounding the effect of prohibited letters (Orth-03). Prohibited letters were overwhelmingly used in the first two decades of the programme (two thirds of them are in names approved before 1960), and their usage has steadily decreased over time (see Appendix section A9). Since historic guidelines are not available, it may be that this rule was not in place during those years, and naming styles will of course have evolved over time. Many of these early names are still prescribed very frequently, such as prednisolone, with over 89 million items prescribed annually in both England and Wales. Both year and prescribing frequency are significantly associated with error, with older names and those with high usage associated with increased risk of error, and this may be drowning

out any effect of prohibited letters. There simply are not enough examples of them in recent years to determine an effect. In fact, in isolation, names containing prohibited letters are involved in more LASA errors than those not containing them (30% with, compared with 23% without) so a protective effect against error is not certain.

The measure is also oversimplified: a binary yes/no for containing any of those letters, with no disaggregation of which prohibited letters are driving the effect. Thus, it may be that certain letters prime for the risk of error but not all. Future research could disaggregate which prohibited letters are problematic because some are homophones but others represent sounds that do not have an alternative single letter grapheme in English and perhaps do not exist in other languages, e.g. <th> for /ð/ or /þ/. Chapter 6 problematises the prohibition of certain letters that actually reduce ambiguity in pronunciation of names, e.g. <k>, which is a useful way to distinguish between hard /k/ and soft /s/, both of which can be represented by <c> in English, e.g. amoxicillin and acetylcysteine. The letter <k> in particular is the prohibited letter that persists in use even in more recent decades, because it forms part of pharmacological STEMS., in particular those for monoclonal antibodies, such as -kinumab in canakinumab. Given its usefulness in translation it seems prudent to review whether this letter should be prohibited in the naming guidelines. Any change to guidelines should be carefully considered, including restricting certain letters – changes could have an impact on orthographic neighbourhood and could introduce new problems such as difficulties in typing, spelling, searching for names (c.f. Ferner & Aronson, 2016). Future research is needed but researchers and naming bodies should note that restricting certain letters is a complex issue.

7.2.3 Morphosemantic features of INNs and their association with LASA error

In Chapter 4, eight morphosemantic features of INNs are explored, many of which explored a subset of the dataset determined by compliance with WHO guidelines (WHO, 2021b), being listed in the WHO Stem Book, and forming part of a STEM taxon with at least five names. Chapter 4 also qualitatively describes the semantic motivation of STEMS and characterises the formation of names. Features analysed under this category are: compliance with WHO naming guidelines (Morph-01); presence of a STEM (MORPH-02), SUBSTEM (Morph-03), variant STEM (Morph-04), incidental STEM (Morph-05), or FREEFIX STEM (Morph-06); headedness (Morph-07); and taxon type (Morph-08) as per the typology defined in Chapter 2, Section 2.4 and from Bryan (2015). The findings for these are summarised in Table 7.2. This thesis addresses a shortcoming in the literature by analysing semantic features of drug names, so much of the findings on morphosemantic features are novel, or

substantively expand the small-sample analysis in Bryan et al. (2015) of 71 names. Later in Chapter 5, Morph-08$_{\text{taxon type}}$ was expanded to include the full dataset by subsuming variables for listedness and compliance, and small taxon size – this is new variable Morph-09$_{\text{STEM taxon}}$. and is included in Table 7.2 below.

INNs are terms that are formed from a "random, fantasy prefix", which in practice may appear anywhere in the name, combined with a STEM and up to two SUBSTEMS (WHO, 2017a). In the dataset of 1,046 INNs, 197 unique STEM families were identified, e.g. -mycin and -astine. STEMS are combining forms that may be neoclassical, such as *chlor* and *phos*, and are ostensibly formed via abbreviating the name of the pharmacological group (e.g. gab for GABamimetic agents), creating a partial acronym of the pharmacological group (e.g. -mab for **M**onoclonal **A**nti**B**odies), abbreviating the chemical name or using a chemical suffix (e.g. -retin for RETINol derivatives), or abbreviating an existing INN (e.g. -azepam for diAZEPAM derivatives). These formation mechanisms were observed in the data rather than being stipulated by the naming body (the WHO).

Findings on morphosemantic features broadly confirm and broaden the findings of Bryan et al. (2015), which analysed a much smaller sample of 71, compared to 1,046 here. Compliance to WHO guidelines is high, and over 80% of names are listed in the Stem Book. Compliance in this sense is prescriptive – it is defined by whether the WHO Stem Book publication deems the name compliant with its guidelines, categorising it as a "name in which the preferred STEM has been used in accordance with its definition". A large number of names in the dataset contain an official FREEFIX STEM – a STEM that can appear anywhere in the name without a change in meaning This is higher than found in previous work (6/43 names) but the current dataset measures a larger sample so provides a more reliable result (Bryan et al., 2015). However, although these STEMS are officially STEMS without a fixed, prescribed position, in practice in the dataset they rarely change position. The majority of names are right-headed, showing their semantic head at the end of the name with a right-to-left display of taxonomy (e.g. -lukast in montelukast). A minority are left-headed, but a significant number are neither-headed. These names display a SUBSTEM via two processes. The first is agglutination (the SUBSTEM is placed at the end of the name, resulting in the STEM [the name's semantic head] being in the middle of the name, e.g. linagliptin with FREEFIX STEM gli and SUBSTEM -ptin). The second is replacement, where the hyponymic SUBSTEM replaces its 'parent' STEM, the semantic head (e.g. tegafur with SUBSTEM -ur replacing 'parent' STEM -uridine).

The problem of incidental STEMS were first introduced in Bryan et al. (2015), along with the concept of a *FREEFIX STEM*. The 2015 paper also noted that some names in that larger dataset of just under 8,000 names incidentally contained FREEFIX STEMS without being members of that STEM taxon, so this was measured for all STEMS in the current study. The majority of INNs contain at least one STEM incidentally but are not a member of that pharmacological STEM taxon. In this dataset there was only one example of a name that incidentally contains a FREEFIX STEM, i.e. a STEM that can appear anywhere in the name and thus communicates meaning regardless of its position – etonogestrel, with official FREEFIX STEM gest (steroids, progestogens), incidentally containing FREEFIX STEM estr (oestrogens). It seems not to be widespread but nevertheless names that incidentally contain a FREEFIX STEM (in whatever position) are more likely to be interpreted as being members of that STEM taxon than those containing other STEMS that have a fixed position when they are used officially. Names containing a STEM without belonging to the STEM taxon are liable to be misinterpreted.

A typology of taxa was first developed in a book chapter (Bryan, 2015) and further elaborated here to include left-headed names. The STEM taxon displayed by the name was classified into one of four types: single-level; regular; complex; and irregular. Since this typology was developed after the 2015 study, the names in the earlier small sample were not categorised, although from the results it can be gleaned that 20 out of 43 names were single-level, 8 were regular, 3 were complex, and the remaining 12 were irregular – the distribution of these numbers is comparable to the results from analysing a much larger sample here. Findings on the morphosemantic features of INNs represent a significant contribution to this field of research, which lacks knowledge in semantic features of INNs. A useful output from characterising the dataset of INNs is a list of STEM taxa, with their respective INNs, collated in three languages as in Appendix section A15.

Table 7.2 Summary of findings on morphosemantic features of INNs

Var	Descriptive findings	Statistical findings	Policy and regulation	Future research
Morph-01 Is the INN compliant with WHO naming guidelines?	80% of names were listed in the Stem Book, and of these, 92% were deemed compliant with WHO guidelines.	Not taken forward for statistical analysis		
Morph-02 Does the INN contain a STEM?	96% contain a STEM, e.g. -mycin in erythromycin.	Not taken forward for		

		statistical analysis		
Morph-03 Does the INN contain a SUBSTEM?	28% contain a SUBSTEM, e.g. -lukast in montelukast.	Not taken forward for statistical analysis		
Morph-04 Does the INN contain a variant STEM?	Only a handful (<2%) of names use a variant form of the STEM, e.g. -ulosin in tamsulosin, under official STEM -azosin.	Not taken forward for statistical analysis		
Morph-05 How many incidental STEMS appear in the INN?	50% of names contain at least one incidental STEM, e.g. -ast in dutasteride.	No		
Morph-06 Does the INN include a FREEFIX STEM?	18% of names contain a FREEFIX STEM, e.g. e.g. gli in sitagliptin.	Not taken forward for statistical analysis		
Morph-07 Is the INN right-headed?	69% of names are right-headed, e.g. atenolol; 2% are left-headed, e.g. cefepime ; 29% are neither, e.g. sitagliptin, with (FREEFIX STEM gli in mid position.	Not taken forward for statistical analysis		
Morph-08[44] Which taxon type does the STEM in this INN belong to? (types 1, 2, 3, or 4)	33% are in a single-level taxon; 20% are in a multi-level taxon; 14% are in a complex taxon; and 33% are in an irregular taxon shape (either not right-headed, or containing a FREEFIX or variant STEM).	Not taken forward for statistical analysis		
Morph-09[45] What are the features of the STEM taxon (listed, compliant, small taxon size, irregular, single-level, multi-level, complex)	19% of names are not listed in the Stem Book; 6% are not compliant with guidelines; 31% are in a small STEM taxon; 14% are in an irregular taxon type; 15% are in a single-level taxon; 9% are in a multi-level taxon, and 6% are in a complex taxon.	As a group p<0.001	There is not enough empirical evidence from the statistical analysis to make any policy recommendations based on this finding.	It appears that high-risk names (in regular STEM taxa) are becoming more prevalent in the nomenclature. Future research could explore the relationship between various characteristics of STEM taxa and error,

[44] Analysed for a subset of names with at least five members in the STEM taxon (n=452)
[45] Calculated for the subset of the dataset analysed in Chapter 5 (1,000 names)

	potentially using other nomenclatures for comparison or by changing the definition of each taxon type to allow more nuance in analysis.

7.2.3.1 Characteristics of a name's STEM taxon (Morph-09)

As an unordered factor, characteristics of the STEM taxon (Morph-09) are found to be highly associated with differences in LASA error risk; see the parameters in Figure 5.5 in Chapter 5. This variable is an expansion of the taxon type variable in Chapter 4 (Morph-08)[46], to include listedness in the WHO Stem Book, compliance, and STEM taxon size. The typology was developed by this author and has not been tested elsewhere, so this piece of analysis further develops and tests it quantitatively and explores its association with LASA error (Bryan, 2015).

The typology is based on certain taxonomy norms, such as those presented by the Office for National Statistics (ONS) and the International Organization for Standardization (ISO). They advise that a robust taxonomy has two key features: internal *coherence*, and *clarity of design*. To achieve *coherence*, relationships and classifications should be logical and consistent, with related concepts grouped together and hierarchical structures following a consistent pattern - this facilitates effective data organisation and retrieval. To achieve *clarity of design*, a taxonomy should be clearly defined and organised so that users can navigate and understand it. This requires clear naming conventions and predictable hierarchical structures to prevent ambiguity and confusion (ONS, 2024). In the typology, three types of taxon are regular and represent increasing complexity; the other 'irregular' type is a 'catch-all' category for features that are not in line with these taxonomy norms. These are STEM-level features of names, rather than individual features, they characterise the STEM that is used in an INN. Not being listed, and not using a STEM in a WHO-compliant fashion, are also features of the pharmacological STEM that may play a role in a user's familiarity and usage of a name, as is whether the STEM is frequently used or only appears in a handful of names.

[46] A typology of taxa was first developed in a book chapter (Bryan, 2015), based on a qualitative analysis of the INN nomenclature analysed in Bryan et al. (2015). Four types of taxon shape, or type, are defined: single-level (a STEM with no hyponymic SUBSTEMS); regular (a STEM with hyponymic SUBSTEMS); complex (a STEM with two hyponymic SUBSTEMS, this type is wholly comprised of monoclonal antibodies named up to 2017); and irregular (FREEFIX STEMS even if they are not mobile, variant STEMS, and STEMS that are not right-headed).

Thus, the combined variable Morph-09 aims to capture overall characteristics of INN STEM taxa.

The findings suggest that irregularity in INNs has a protective effect against LASA error – alone, presence of features such as SUBSTEMS, FREEFIX STEMS, variant STEMS, and non-right-headed STEMS do not have a significant effect, but when they are combined and compared against names without those features an effect emerges. Regular taxon types, therefore, are associated with higher risk of LASA error compared with irregular taxon types and compared with names in small STEM taxa (<five members). It may be that regularity of taxon type drives error by not prompting the attention of users, or by being too complicated so users don't understand them. However, the odds ratios for these groups' association with error have wide confidence intervals and bootstrapping revealed bias in the 'complex' type[47]. The findings do not have strong enough empirical rationale for a policy change, but the picture suggests that trends of each type are changing – the high risk categories are increasing and the lower risk categories are decreasing (single- and multi-level STEMS are being used more frequently and the overall size of STEM families is increasing, see Appendix section A12). This association certainly merits further research to disentangle the effect, to avoid priming the nomenclature for LASA error.

7.2.4 Relational features of INNs and their association with LASA error

Eleven relational features of INNs were characterised in Chapter 4: frequency of initial and terminal bigrams and trigrams; STEM frequency; ATC alignment to two and four levels; neighbourhood density, neighbourhood frequency, and top-prescribed neighbour; and semantic transparency. The endings of INNs are more homogeneous than the beginnings, with over three quarters of names in the dataset ending in the top ten bigrams, such as mercaptamine with <ne>. The "random, fantasy prefix" (WHO, 2017) in the majority of names goes some way to increase diversity in the beginnings of names. At the same time, most STEMS being SUFFIXES reduces diversity in endings as they are constrained by the form of the STEM. The diversity of USAN beginnings and endings is comparable to that of INNs (Lambert et al., 2001a) despite Lambert's study using a larger sample size - the most common ending in USANs is <ine>, in 22.34% (1191/5331 USANs), compared with 21.80% in this dataset (228/1,046 INNs).

[47] Note that the complex taxon type of names for monoclonal antibodies no longer exists, and no other names have been assigned a double SUBSTEM, so in future, regular STEM taxa will be either single- or multi-level.

Alignment between the INN and ATC taxonomies is relatively low. When explored for the top two levels of ATC classification (the broad, general conceptual categories) the two systems aligned well. This was measured by the extent to which names in a STEM taxon share the first two levels of the ATC code, e.g. the L04 of infliximab. However, when all four levels in the ATC classification hierarchy were explored, it was not well aligned with the INN taxonomy. This measured the extent to which they share all four levels of ATC code, e.g. L04AB of infliximab) was negative. So there is alignment in top-level categories but divergence when the concepts become more specific. This may reduce predictability if users assume mutual sense between the two systems (because they are both administered by the WHO). Overall, semantic transparency (measured by the extent to which a name's semantic head is predictable within its STEM taxon) is deemed high across the dataset, with little movement of FREEFIX STEMS in practice, and a preponderance of right-headed names using STEMS from regular taxa. Internal consistency within STEM taxa enhances predictability of meaning (i.e. by signifying membership of the taxon), but misalignment between ATC and INN classification systems may hinder user understanding.

Most names in the English-language dataset have between one and five neighbours and some have no neighbours. Names with a large neighbourhood are observed to use frequently-used STEMS such as -mab and -tinib. Just under half of names have a neighbour that is prescribed more frequently than itself, and some have a super-prescribed neighbour. Neighbourhood density has not been measured before in INNs, though it has been measured in USANs by Lambert et al. (2001a) and the results are comparable, despite measuring different generic nomenclatures and being two decades apart. In this dataset of INNs, 0.44% pairs of names (4,347/1,093,070 pairs scored) were over the imposed similarity threshold of 33%, using the trigram-2b score. Lambert et al. found that 0.28% (47/16,641 pairs scored) of USANs were between 30% and 40% similar using the same score (Lambert et al., 2001a). The findings on neighbourhood density represent a significant contribution to the literature on orthographic similarity between drug names, and to date this is the biggest sample to have been scored. The finding that names with a high number of neighbours tend to use common STEMS aligns with findings from previous work that revealed a tension between competing goals of reducing similarity while displaying semantic relationship. In other words, using STEMS to show pharmacological relationships increases internal pairwise similarity in the nomenclature (Bryan et al., 2015) and this analysis confirms this finding.

Of the eight relational variables brought forward for statistical analysis in Chapter 5, five remained in the final model, and are discussed in turn in the subsections below. These are:

initial trigram frequency (Rel-02); terminal trigram frequency (Rel-04); semantic transparency (Rel-10); neighbourhood frequency (Rel-09); and ATC alignment (Rel-07).

Table 7.3 Summary of findings on relational features of INNs

Var	Descriptive findings	Statistical findings	Policy	Future research
Rel-01 How frequently do the first two letters of the name (initial bigram) appear in the data set?	Mean frequency of initial bigrams is 12.96.	Not taken forward for statistical analysis		
Rel-02 How frequently do the first three letters of the name (initial trigram) appear in the data set?	Mean frequency of initial trigrams is 3.75.	p<0.001, OR 1.081 [1.038, 1.126] This findings confirms association found other studies. It also supports the practice of weighting the beginning of the name in orthographic similarity measures, to emphasise the importance of the initial string.	A tentative suggestion that manufacturers proposing new INNs and the INN Expert Committee assessing proposals encourage diversity in the beginnings of new INNs – lists of frequent initial strings are given in Table 4.14 in Chapter 4.	
Rel-03 How frequently do the last two letters of the name (terminal bigram) appear in the data set?	Mean frequency of terminal bigrams is 133.63.	Not taken forward for statistical analysis		
Rel-04 How frequently do the last three letters of the name (terminal trigram) appear in the data set?	Mean frequency of terminal trigrams is 72.21.	p=0.017 OR 1.002 [1.000, 1.004] This confirms association found other studies, however the association is overinflated (relative conditional bias of 40%) and it is present in fewer models than Rel-02 (81%).	The finding on an association between *terminal* trigrams and error is not as robust as the finding on *initial* trigrams. However, discouraging the use of -ine in new stems would halt the dominance of this ending in the nomenclature. Names ending in -ine make up 22% of the dataset but 31% of LASA errors. Lists of frequent initial strings are given in Table 4.15 in Chapter 4.	Given the finding of terminal trigrams (Rel-04), there is potential in weighting the name at the end as well as the beginning, so future research could explore other weighting approaches, such as 'trigram-2b-1a'.

Rel-05 How frequently does this INN's STEM appear in the dataset?		No		
Rel-06 To what extent do the ATC codes assigned to the INN align with the ATC codes for other names in the same taxon? (top two levels of ATC)	Some names were well aligned with the ATC system to the top two categories (mean score 0.04, where -1 signifies completely different, and 1 signifies identical).	Not taken forward for statistical analysis		
Rel-07 To what extent do the ATC codes assigned to the INN align with the ATC codes for other names in the same taxon? (all four levels of ATC)	Less alignment to all four categories of ATC (mean score 0.24 – a lack of alignment with the ATC system.	p=0.096 OR 1.712 [0.898, 3.205] Despite a positive effect, the confidence interval is large and spans 1 so the direction of the effect is indeterminable. It is not significant in the global model and is not significant to <5% in the final model (p=0.096), it is heavily biased upwards (54%), and is in just over half of the bootstrapped samples.	There is not a strong empirical rationale to recommend any policy changes based on this finding.	
Rel-08 Neighbourhood density: how many names are similar to the INN?	58% have between one and five neighbours and 13% have no neighbours	No		
Rel-09 Neighbourhood frequency: are any of the INN's neighbours prescribed more frequently?	42% of names have a neighbour that is prescribed more frequently than itself	p=0.081 OR 1.367 [0.963, 1.943] This association is not certain: the p value is significant to 10% but not to 5%, the confidence interval is broad and spans 1, the variable is in only 64% of bootstrapped resamples and the effect is biased upwards (29%).	There is not a strong enough empirical rationale to recommend a policy change based on this finding. A stronger association would warrant including an assessment of neighbourhood frequency in the INN Committee's assessment of similarity using POCA.	Inferences of the results merit further research into the association between neighbourhood frequency and LASA error, assessing methods to calculate neighbourhood frequency and exploring the feasibility of integrating neighbourhood frequency into the INN decision-making process.
Rel-10 Semantic transparency,	Many names are semantically	p=0.030 OR 0.496 [0.261, 0.930]	Findings here suggest that semantic	The association between semantic

measured by positional family size ratio: to what extent is the INN's STEM in a typical position for names in the same taxon?	transparent (mean score 0.35, where -1 signifies completely opaque, and 1 signifies completely transparent)	This association is tenuous: although the p value was under 5% in both the global and final models, the variable is in 73% of the bootstrapped resamples and effect is biased upwards (relative conditional bias: 37%).	transparency of names is steadily increasing over time. I would tentatively suggest that manufacturers proposing new INNs note the STEM positioning of existing names in that pharmacological group, to support the positive trend.	transparency and LASA merits further research, perhaps in other nomenclatures and other languages.
Rel-11 Are any of the name's neighbours in the top 25 most prescribed names?	11% have a super-prescribed neighbour, such as remifentanil and highly-prescribed fentanyl	No		

7.2.4.1 Beginnings and endings of names (Rel-02 and Rel-04)

Having commonly-used initial and terminal trigrams (three-letter strings) is strongly associated with an increased risk of LASA error. The reason for this is simple: a lack of diversity in how names begin and end means that pairs of names are more likely to share (orthographic) features and thus be confused with each other. The most frequently-used initial trigrams are: flu-, pro-, car-, met-, chl-, clo-, tri-, cef-, lev-, hyd-, ben-, tra- and the top terminal trigrams are: -ine, -ide, -one, -mab, -nib, -cin, -vir, -ole, -tin, -tan. The findings of this study tentatively suggest that those manufacturers proposing INNs and the WHO Expert Committee approving INNs should consider the importance of diversity in the letters at the beginnings and endings of names.

Rel-02$_{\text{initial trigrams}}$ is strongly statistically significant in both the global and final models and is in 97.3% of all bootstrapped models. Given the range is 19, the effect size of Rel-02$_{\text{initial trigrams}}$ means that there is a four-fold increase in odds of LASA error between the least common initial trigram (e.g. avi in avibactam) and the most common (e.g. flu in flupentixol) – the risk of LASA error increases with the frequency of the initial string. Rel-04$_{\text{terminal trigrams}}$ is also statistically significant in both the global and final models. However, the importance of Rel-04$_{\text{terminal trigrams}}$ is inflated in the model, with a relative conditional bias of 40% and present in 81% of bootstrapped models. The effect of this variable is likely dominated by a single very frequent name ending: -ine. Names ending in -ine make up a fifth of the dataset but almost a third of errors, so reducing the use of this ending in pharmacological STEMS and more generally in names is a feasible way to reduce LASA error risk.

This association confirms findings in the literature but uses a different methodology to those studies. Both Shah et al. (2017) and Thompson (2008) describe features of pairs of names that have been reported in LASA errors and found that initial strings in particular featured prominently in LASA error pairs, followed by terminal strings. The findings here on the importance of the initial trigram confirm those from these earlier descriptive analyses and support the approach of weighting the beginnings of names in orthographic similarity measures (Lambert, 1997; Lambert et al., 1999). Lambert et al. (Lambert et al., 1999) have previously tested a 'trigram-2b-2a' score, adding two spaces before and two spaces after the name, but it did not perform as well as 'trigram2b' with just two spaces before the name. However 'trigram-2b-1a' performed similarly well to trigram2b, so there is potential in weighting the name at the end and this is a worthy avenue of future research.

Tables 4.14 and 4.15 in Chapter 4 giving frequently-used strings are of use to manufacturers and naming bodies for avoiding future homogeneity and are useful outputs. Most STEMS are at the end of names, with the beginning advised to be a "random, fantasy prefix" (SoINN, 2020b; WHO, 2017a) so it is easy to increase the diversity of initial trigrams. However, since the name is restricted by the form of the STEM, increasing diversity in terminal trigrams presents more of a challenge. This falls under the remit of those creating new STEMS, in this case the INN expert committee. Diversity would be increased by avoiding further use of the preponderant <ine> at the end of several STEMS, and in general working towards a range of STEMS with novel letter strings. The split taxon for monoclonal antibodies into four novel STEMS is a good example of this: -tug, -bart, -mig, and -ment (Balocco et al., 2022; World Health Organization, 2022).

Over time, the average frequency of both Rel-02 and Rel-04 is falling, meaning that diversity is increasing in both beginnings and endings (see Appendix section A12, *relational measures over time*). However, names using high frequency strings continue to be approved in recent decades, such as carfilzomib in 2008 (20,000 items prescribed annually in England and Wales) and vortioxetine in 2012 (320,000 items prescribed annually).

7.2.4.2 Semantic transparency, and ATC alignment

Semantic transparency (Rel-10) of the name is significantly associated with LASA error. However, the association is tenuous and the effect is biased upwards. Relatively high semantic transparency, attained in this case by displaying the STEM in a predictable position in the name that is consistent for all names within a particular STEM taxon, is ostensibly

found to be associated with a lower LASA error rate. Semantic transparency is a measure of the internal consistency of a STEM taxon. Semantic transparency is correlated with ATC alignment (Rel-07), which is also in the final model, indicating that they share information – it seems that *uniformity* within the taxon plays a role with how the name is recognised, remembered, and differentiated. The finding for ATC alignment (Rel-07), however, is revealed to be very unreliable by the stability investigations in Chapter 5, and the association is likely related to the uniformity of the stem taxon than ATC codes specifically. Despite the metrics of the stability investigations, the relationship between semantic transparency and LASA error merits further research.

For example, all members of the STEM taxon -conazole (systemic antifungal agents, miconazole derivatives) place their STEM at the end of the name thus are deemed semantically transparent. The taxon has a low LASA error rate (10%), e.g. ketoconazole and fluconazole. In contrast, members of the STEM taxon -azepam (diazepam derivatives) are deemed semantically opaque, with five variant forms of the STEM (e.g. diazepam, midazolam, nefopam, clobazam)[48]. The -azepam taxon has a high error rate (60%): e.g., diazepam and lorazepam.

In the previous section the association between STEM taxon characteristics (Morph-09) and LASA error was discussed. A tentative finding from Chapter 5 was that names with an irregular STEM taxon type were associated with reduced rates of LASA error. At first it seems contradictory that both semantic transparency of names and irregularity of STEMS should protect against error. Both pertain to the predictability of meaning but in different ways. It seems that it is useful to have complexity in the nomenclature as long as there is consistency with the STEM taxon - the system can cope with inter-taxon diversity, but not intra-taxon diversity. Uniqueness in how STEMS behave, such as being left-headed or FREEFIX, or having a variant spelling, may prompt users to pay more attention to the name and note its difference from other names. This finding gives further weight to increasing the diversity in initial and terminal trigrams. This does not mean that all names should display their STEM at the end of the name, or the beginning, just that the position should be consistently applied for all members of the taxon. At first glance this implies that FREEFIXES are drivers of error, but in practice most FREEFIXES are in fact not mobile - they generally

[48] Variant forms of the STEM were included in this variable as names that do not apply the STEM in a consistent way, to account for a lack of predictability when using a variant STEM. However, variant STEMS only appear in the dataset in 13 names so this is not a widespread issue in the nomenclature.

appear in the same position within a taxon with a handful of exceptions. For example, the FREEFIX STEM vir appears in 41 names in the data set; in 33 of these it is at the end of the name and in only eight names does the STEM appear in the middle of the name (e.g. maraviroc and doravirine).

Earlier work posited that increased semantic transparency (in that particular case, by showing relationship using a common STEM) also increases orthographic similarity, which then increases the risk of error (Bryan et al., 2015). This is not the case according to the results of this analysis. It seems to be beneficial to have internal consistency of position within a STEM taxon regardless of where the STEM is placed. These findings likely contrast with earlier work because this study has a much bigger sample size and actually quantifies semantic transparency, rather than simply hypothesising an association based on an observation of shared STEMS and orthographic similarity. Despite contrasting with earlier work, this finding aligns with the general consensus that transparency increases processing speed by encouraging 'expectedness in semantic structure' (Bell & Schäfer, 2016). The present analysis focusses on terms in a closed nomenclature that are partial neoclassical compounds, and as such complements the body of literature on semantic transparency, which predominantly analyses general language compounds (Feldman et al., 2004; Libben, 1998; Libben et al., 2003). The present analysis of semantic transparency is a critical expansion of Bryan et al. (2015) and includes a quantification of semantic transparency in a dataset of generic drug names (this is a novel output). It measures semantic transparency using positional family size ratio, a measure first developed by Bell & Schäfer (2016) and applies it to a terminological domain, in this case pharmaceutical nomenclature.

In essence, maintaining a consistent placement of the STEM within its taxon increases semantic transparency, which may in turn reduce the risk of LASA errors. This does not lead to the suggestion to place all STEMS at either the beginning or end of a name; rather, it tentatively suggests that it is important to uniformly apply the position of STEMS across all members of the taxon. In practice, manufacturers proposing INNs are able to use the Stem Book to note the STEM positioning in other members of the proposed name's pharmacological group. Of course, further research is recommended to explore this association, and currently there is not a strong enough empirical rationale to suggest a policy change. It is encouraging that over time, semantic transparency is steadily increasing, from an average of 0.13 (low transparency) pre-1970, to an average of 0.70 (high transparency) post-2000. This means that potential risk is being reduced through consistent use of STEMS in new INNs.

7.2.4.3 Neighbourhood frequency

Neighbourhood frequency, i.e. in the sense of a name having a neighbour that is prescribed more frequently than itself, was found to increase the risk of LASA error, though it should be noted that neighbourhood density was not significantly associated. The association tentatively suggests that neighbours with higher prescribing frequency are likely to 'drown out' a name and increase the possibility of error. This can be mitigated by an overall reduction in the size of the name's neighbourhood, for example by using diverse beginning and ending strings, where STEMS allow. However, the confidence intervals are broad so this result is tentative and warrants further exploration. The stability investigations also demonstrate that the influence of this variable is overinflated and so its association with error is difficult to determine.

Additionally, this finding contrasts with the findings in Lambert, Chang, & Gupta (2003), who found that neighbourhood frequency had an interactional effect with stimulus frequency (the prescribing frequency of a name) and neighbourhood density, but alone was not significantly associated with visual perception. However, the study by Lambert et al. does not measure association with error, just with visual perception of drug names. It also calculates neighbourhood frequency differently, with the mean frequency of all neighbours, whereas the original method from Andrews (1997) is used, which is a binary score of whether any neighbour is more frequent than the name.

However, neighbourhood frequency is easily quantifiable using a range of useful measures (Lambert et al. 1997). Currently, the INN Expert Committee uses a similarity measure created by a Committee member, Dr Andrew Martin, to assess proposed names for monoclonal antibodies but relies on collective memory and *ad hoc* database searches of publishes names to identify similarities (personal communication, November 2022). While the POCA (Phonetic and Orthographic Computer Analysis) score has been referenced in documentation, the Committee found it to be unreliable and subsequently abandoned its use (relayed from email communication with Dr Jeffrey Aronson; for information on POCA, see US FDA, 2007; SoINN, 2020a). It remains to be seen if it would be possible to include a consideration of the prescribing frequency of a proposed name's neighbourhood into the Committee's process. Neighbourhood frequency remains steady over time so is a persistent issue (see Appendix section A12). Further research is required on the association between neighbourhood frequency and LASA error, assessing methods to calculate neighbourhood frequency and exploring the feasibility of integrating neighbourhood frequency into the INN decision-making process.

7.2.5 Summary of findings for Research Questions 1 and 2

Chapter 4 describes the creation and features of a taxonomic database of INNs, answering Research Question 1. It expands Bryan et al. (2015) and complements Lambert et al. (2001a). Results were mostly confirmatory, aligning with findings for the smaller sample size in Bryan et al. (2015) and findings on USANs in Lambert et al. (2001a). This characterisation represents several significant contributions to existing research on drug nomenclature, bringing new insights into orthographic features such as Scrabble score, morphosemantic features such as the semantic motivation of STEMS, and a comprehensive taxonomic classification of INN STEM taxa. Other new knowledge and insights about INNs include semantic transparency, alignment with ATC classification, and neighbourhood measures. Overall, these outputs contribute to a deeper understanding of the INN nomenclature and provide a useful framework for analysis of other generic name systems or generic names in other languages.

Chapter 5 describes the statistical analysis of association between linguistic features of names and the risk of being involved in a LASA error, working to identify linguistic latent conditions that prime the risk of LASA error. This is in answer to Research Question 2. Findings are discussed in terms of tentative *suggestions* and *considerations* for manufacturers of INNs and the INN Expert Committee, and where appropriate directions for future research are highlighted.

Four *a posteriori* assumptions of connected features are proposed, three of which were confirmed: high-frequency initial and terminal strings, and neighbourhood frequency (neighbourhood density was not). With the caveats discussed earlier in this section, the findings broadly confirm what is found in the literature on features associated with LASA error but use complementary methodologies. For example, Lambert et al. (2003) found an association between neighbourhood (density and frequency) and visual perception of drug names – here association with LASA error is measured using real-world error data. Shah et al. (2017) found an association between high-frequency initial and terminal strings from a descriptive analysis of error data – here regression techniques are used to measure the association.

Overall, the statistical analysis creates several useful contributions to the literature on linguistic features associated with LASA error. By pinpointing associated features, knowledge created in characterising the nomenclature in Chapter 4 can be used to inform naming bodies on features to avoid or promote in the creation of future names. As examples,

lists of high-frequency initial and terminal trigrams in Tables 4.14 and 4.15 can be used to increase diversity in the beginnings and endings of names, and the typology of taxa developed further in Chapter 4 is a useful framework to assess the complexity of STEM taxa.

The statistical analysis identified a strong association between initial trigram frequency (Rel-02) and LASA error, and makes the suggestion that suggest that those manufacturers proposing INNs and the WHO Expert Committee approving INNs should consider the importance of diversity in the letters at the beginnings and endings of names. The study also identified associations between six other linguistic features and error, but in these cases the effect is unclear and unreliable when subjected to the scrutiny of stability investigations, using bootstrapping. These were: STEM taxon characteristics (Morph-09); terminal trigram frequency (Rel-04): semantic transparency (Rel-10); presence of WHO-prohibited letters (Orth-03); neighbourhood frequency (Rel-09); and ATC alignment (Rel-07).

The significance of the two non-linguistic confounders – prescribing frequency and age of the name – cannot be ignored in any discussion of LASA error. Experimentally, prescribing frequency is correlated with a reduced risk of LASA error, i.e. lesser-used names are less likely to be confused. This is not based on LASA error reports, but rather on a lab-based visual perception task (Lambert et al. 2003). However, those authors acknowledge that despite finding that increased prescribing frequency reduces error, in practice the number of actual errors is a function of opportunity and probability of error: low-risk names with high prescribing frequency may result in more errors than high-risk names with low prescribing frequency, simply because the usage of these names is of a magnitude higher. This statistical analysis confirms the position of Lambert et al. (2003). Organisations managing LASA error watch lists such as the ISMP should take these factors into account when creating lists of potential LASA name pairs.

7.3 Research Question 3: Findings

Drug nomenclatures in two languages are characterised in Chapter 6, adopting a case study approach. These nomenclatures are Russian-language INNs (designed by the WHO) and Italian Common Names (designated by the Italian Pharmacopoeia). Implications of findings are then considered from the statistical analysis (in Chapter 5) in these nomenclatures, theorising how linguistic latent conditions manifest in translation and/or transliteration into Russian and Italian. Chapter 8 answers Research Question 3: *How can Russian-language*

and Italian-language pharmaceutical nomenclatures be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?

According to WHO publications, English is used as the *de facto* source language for translation of INNs into Russian (and the other official UN languages), but it is unclear which language is used as the source language for translation into Italian. Latin and French are mentioned in the literature as potential source languages (Marecková & Cerveny, 2002).

7.4.1 Russian-language INNs: findings from the first case study

Russian names are a little shorter than English names, score higher in Russian Scrabble than English names do in English Scrabble, and use a smaller alphabet than the general language of 26 letters (rather than 33). The excluded Russian letters represent sounds that do not have an English equivalent, e.g. <щ>, for /ɕː/, close to 'shcha' as in *Welsh sheep*. Russian names closely follow the ISO 1995 transliteration standard and the phonological form of the name is transliterated, resulting in the loss of final non-syllabic <e> in number of names, e.g. mitoxantrone > митоксантрон, while keeping terminal vowels that *are* syllabic, e.g. carbidopa > карбидопа. Prohibited letters are treated phonemically, transliterating both <y> and <i> as <и>, as in demeclocycline > демеклоциклин. The prohibited letter <k> has a close equivalent in Russian and its use in English is useful for disambiguating hard /k/ and soft /s/, at times both represented by <c> in English, as in vernakalant > вернакалант.

Information on STEMS is only available in English for English-language INNs, but the Russian case study presents a qualitative analysis of *assumed STEMS* in Russian names, drawn from an observation of translated forms. Russian STEMS are in the same position in their names as their equivalent English forms, even when STEMS are FREEFIXES – there is no evidence that names are decomposed in translation beyond the presence of morphs that already exist in the scientific domain of the Russian language, such as *meth* and *chlor*, as in chloroxylenol > хлороксиленол. Headedness does not change in translation. Right-headedness is less dominant in Russian general language as is also the case for English, but these are partial neoclassical compounds which are typically right-headed, so the difference is less stark.

For the most part, the semantic motivation of English STEMS is maintained in Russian, in particular for STEMS that abbreviate the pharmacological group or chemical name (e.g. -dil for vasodilators, in Russian ***вазодилататоры*** [pron. 'vazodilatori'], e.g. minoxidil > миноксидил) or abbreviate an existing INN (e.g. ketoconazole > кетоконазол). STEMS that are formed via a partial acronym of the pharmacological group are decoupled from their

semantic motivation in translation, due to a change in word order or the partial use of native forms rather than neoclassical forms (c.f. -mab, monoclonal antibody > -маб, моноклональное антитело, which would form the initialism -мат. The Russian language features native forms alongside neoclassical combining forms, e.g. *гидротерапия* 'gidro', cognate with 'hydro', and водопроводод, 'voda', native form of water. When native forms are used instead of neoclassicals in either the source or the target language the semantic links are broken (Georgieva, 2013; Iacobini, 2004). This is purely theorised here since the only primary data in Russian used in the dataset is the Russian name itself, as no information on Russian STEMS was available.

There is significant crossover between initial trigrams in Russian and English due to close transliteration, but there is more divergence in terminal trigrams, due to the final letter being dropped from English in many instances. Neighbourhood density was marginally higher within the Russian dataset than in English with 60% of Russian names having between one and five neighbours. Phonemic transliteration and the simplification of homophonic letters in English (such as <ph> and <f>, <y> and <i>) causes an increase in orthographic similarity, which in turn increases neighbourhood size.

7.4.2 Italian-language DCIts: findings from the case study

Italian names are a little longer than English names, score higher in Italian Scrabble than English names do in English Scrabble, and extend the Italian alphabet with three borrowed letters: <k>, <w> and <x>. Almost half of Italian names are identical to their English counterparts (e.g. ifosfamide > ifosfamide), and where they do change, this is typically to add a vowel to a terminal consonant (e.g. argipressin > argipressina) or to change the terminal vowel (midodrine > midodrina; salbutamol > salbutamolo). However, even those names that do not undergo an orthographic change can undergo a phonological change, because final non-syllabic <e> in English becomes syllabic in Italian, e.g. sertraline > sertralina, and some graphemes are pronounced differently in the languages, e.g. <z>, pronounced /ts/ in Italian, e.g. in pantoprazole > pantoprazolo. Some consonant clusters that are not part of standard Italian phonotactics are domesticated, e.g. <ct> to <tt> in lactulose > lattulosio. As in Russian, homogenisation of name endings can correct meaningless variation in STEMS, e.g. -ifene and variant STEM -ifen. Although Italian names are not subject to WHO naming guidelines, the way in which prohibited letters are treated in translation is explored. For the most part they are treated phonemically, rendering both <y> and <i> as <i>, as in benzydamine > benzidamina and approximating phonemes that do not exist in Italian, such as rendering <th> (for ð and θ) as <t>.

As in the Russian case study, information on STEMS is only available in English for English-language INNs, so instead the Italian case study presents a qualitative analysis of *assumed* STEMS in Italian names, drawn from an observation of translated forms. Italian STEMS are in the same position in their names as their equivalent English forms, even when STEMS are FREEFIXES. Names do not seem to be decomposed in translation beyond morphs that already exist in the scientific domain of Italian, such as *meth* and *chlor*, as in sulfamethoxazole > sulfametoxazolo. Headedness does not change in translation, and the nomenclature is still overwhelmingly right-headed. Italian is a left-headed language, although neoclassical forms are typically right-headed, so it is unclear how 'foreign' the difference in headedness would feel to a native Italian speaker.

As in Russian names, the semantic motivation is maintained in Italian for the most part, but this is problematic for those STEMS that are formed via partial acronymisation, such as -mab for monoclonal antibodies, which would be *-acm in Italian, for *anticorpi monoclonali*, due to differences in word order and the use of non-classical combining forms in one of the languages (in this case, English *body*, c.f. Iacobini 2004).

There is a lot of convergence of frequent initial trigrams between English and Italian, but terminal strings differ because of changes in the last letter, either adding a vowel or changing the terminal vowel. Italian name endings are more homogeneous than those in English, because English names ending in <in> and <ine> are similarly translated as <ina> in Italian, e.g. verteporfin > verteporfina; naphazoline > nafazolina). Neighbourhood density distribution was different in the Italian dataset, with a lower proportion of names with zero or ≤5 neighbours, but more names having super dense neighbourhoods with up to 30 neighbours (e.g. axitinib, with afatinib, upadacitinib, etc.). As in Russian names, phonemic translation and the aggregation of homophonic letters in English (such as <ph> and <f>, <y> and <i>) causes an increase in orthographic similarity, which in turn increases neighbourhood size.

7.4.3 Implications of findings from RQ2 for Russian and Italian drug names and their (potential) association with LASA error

The implications of findings on linguistic features potentially associated with LASA error were explored within the context of Russian- and Italian-language nomenclatures, with some speculations. First, increased diversity in Russian terminal trigrams may amplify the protective effect against LASA error (and conversely, reduced diversity in Italian terminal

trigrams may dampen the protective effect). This is a result of vowels being dropped in Russian and being added in Italian, contrasted with very little change to the beginnings of names. Second, the protective 'jarring' effect of a difference in headedness in English (from right to left, or endocentric to exocentric, compounds) may be lost in Italian, and to a lesser extent in Russian, because right-headedness is not the dominant form in these languages. Third, Italian is a left-headed language with right-headed forms mostly restricted to neoclassical words and loanwords, so it is possible that semantic transparency is lower because users will not parse names in the expected right-to-left order, though this is likely counteracted by INNs' formation as partial neoclassical compounds. Fourth, longer names may have a protective effect on error, though this finding is tentative - if this is the case, this effect would be strengthened in Italian as the names are longer on average, and dampened in Russian as the names are shorter. Fifth, another tentative finding is the protective effect of the presence of prohibited letters. In using a restricted alphabet the Russian nomenclature may feel more orthographically shallow to speakers, which in turn weakens the protective effect, and vice versa for Italian, which uses an extended alphabet. Sixth, neighbourhood frequency may play a role in driving errors involving Russian and Italian names because it interacts with neighbourhood density, which is higher in these languages than in English. It must be noted, however, that these observations are speculative and have not been tested due to a lack of data on LASA errors in both Russian and Italian.

7.4.4 Summary of findings for Research Question 3

Findings from the two case studies reveal interesting parallels between nomenclatures in Russian and Italian that transcend their differences. They are both close in orthographic form to the English-language INNs, with Russian complying closely with the international transliteration standard and many Italian names being identical to English names. Phonological changes were outside the scope of analysis here but changes (or lack thereof) in orthography will result in phonological change, which warrants further exploration.

The translation of International Nonproprietary Names (INNs) into other generic name systems is a form of morphological calquing, where the form of a STEM remains consistent within its STEM taxon (e.g., -cillin consistently translates to -циллин[49] in Russian and -cillina in Italian, as in ampicillin > ампициллин; ampicillina). Existing neoclassical and scientific forms of morphs are used where possible, and this is systematically applied, e.g. both *chlor* and *hydro* in hydrochlorothiazide > гидрохлоротиазид; idroclorotiazide. The "random,

[49] Pronounced /tsɨˈlʲin/

fantasy" part of the name is also calqued into the target language. Loss of semantic motivation for STEMS is complicated in the same way in Russian and Italian, by the presence of native equivalents of neoclassical combining forms in the English language, such as *corpus* and *body* in English, and in Russian (e.g. ***гидро*** and ***вода***) that are not seen to the same extent in Italian (Georgieva, 2013; Iacobini, 2004; Puato, 2011, 2013).

The translation strategy of morphological calquing, using existing morphs where possible, translating STEMS consistently and maintaining their position in the word, while meeting the phonotactics of the target language, aligns with Nida's theory of *formal equivalence*. Applied to drug nomenclature, this theory emphasises maintaining the form and terminology of the source text as closely as possible while ensuring clarity and understanding in the target language (Nida, 1977; Statham, 2003).

The similarity of features between Russian and Italian nomenclatures is interesting because Russian names are international names approved by the WHO INN committee, whereas Italian names are national and approved by a national pharmacopoeia. This alignment between two disparate systems points to a level of standardisation that goes beyond national and organisational boundaries, implying that certain translation conventions prevail regardless of the nomenclature's origin or regulatory context.

Chapter 8 provides several useful contributions to the literature, including the first analyses of linguistic features of generic drug names in Russian and Italian languages, and presentation of STEM taxa in all three languages including English. Table A8.1 in Appendix 8 is a valuable list of INN STEMS in all three languages and of names using each STEM.

7.4 Conclusion

The goal of this thesis is to identify linguistic latent conditions in international drug names and understand the implications of findings for nomenclatures in Russian and Italian. To achieve this aim, three research questions (RQ) were asked: *RQ1, How can the International Nonproprietary Name (INN) nomenclature be characterised in linguistic terms?*; *RQ2, To what extent are any of the linguistic features described in answer to RQ1 associated with the occurrence of LASA error?*; and *RQ3*, *How can Russian-language and Italian-language pharmaceutical nomenclature be characterised in linguistic terms, and what are the implications of findings from RQ2 for these languages?*

LASA errors are a type of medication error between similar drug names and constitute a significant threat to patient safety. They are highlighted in the WHO's 2017 global patient safety challenge *Medication without harm* (WHO 2017). Linguistic features that increase the risk of LASA errors are framed as a type of latent condition in the healthcare system that creates 'accident opportunities', under James Reason's human error theory. Linguistic features have been explored to an extent in the literature, but there is a lack of knowledge on morphosemantic features of drug nomenclature (most studies look at formal features such as name length, e.g. Lambert et al. 2001a), linguistic properties of generic names (most look at brand names, e.g. Basco et al., 2010, Shah et al., 2017), and drug nomenclature in languages other than English.

In conclusion to Research Question 2, one **linguistic latent condition** is identified - initial trigram frequency – which leads to a suggestion for manufacturers and regulators to consider encouraging diversity in the beginning strings of names. Several other linguistic variables show potential associations with LASA error that merit further research but fall short of sufficient empirical evidence to make any policy recommendations. This is the first study to use statistical methods to identify morphosemantic factors that drive error. Furthermore, a strongly significant association of two non-linguistic confounders is identified: prescribing frequency and year of approval. This is a novel output in the field and contributes to the policy-facing work in LASA error reduction.

I have proposed to present these findings at the INN Stakeholder session in April 2025 (I presented preliminary results in October 2022), and publications are in preparation for PlosOne for a 'results' paper, and the British Medical Journal for an 'overview of INNs for clinicians', following discussions with editors.

7.4.1 Contributions to the literature

This study addresses three notable shortcomings in the literature, namely lack of knowledge on morphosemantic features of drug names, linguistic features of generic names, and drug nomenclatures in languages other than English. Overall I have adopted a whole-system approach, exploring drivers and interventions in two peer-reviewed literature reviews (Bryan et al., 2020a, 2020b).

In Chapter 4, a detailed characterisation is presented of the orthographic, morphosemantic, and relational features of INNs (N=1,046), which are the globally-preferred form of generic drug names. The morphosemantic analysis explored in depth morphological composition,

semantic motivation, semantic relationships—including hyponymy—and the taxonomy represented by STEMS and SUBSTEMS. This expands and validates the small-sample study led by this author in Bryan et al. (2015). In analysing a basket of linguistic features in three languages – English, Russian, and Italian – INNs are situated within the multilingual and international system of drug nomenclatures, in this case regulated by the World Health Organization (WHO) and the European Medicines Agency (EMA). In many respects Russian INNs and Italian Common Names are very similar to each other, and to English INNs - this is of particular interest because while Russian INNs are designated by the WHO, Italian Common Names are designated by Italy's national pharmacopoeia (Ufficiale Farmacopea Italiana). Formal features of INNs were similar to those of USANs, complementing Lambert et al. (2001a).

The findings in this study are drawn from several novel analytical components, including the first analyses of semantic transparency and taxon type in English INNs (and indeed, of any kind of drug name). Both features were found to be potential risk factors for LASA error, and the analysis helped to finetune the taxon typology first proposed in Bryan (2015). The contrastive analyses of Scrabble score, neighbourhood density, beginning and end strings between English, Russian, and Italian are also novel.

In Chapter 5 the association between linguistic features of English-language INNs and the risk of LASA errors are explored using inferential statistics. This builds upon observational findings of associated features in Shah et al. (2017) and Thompson (2008). Practical outputs are produced, such as lists of pharmacological STEMS in English, Russian, and Italian (section A16 in the Appendix), along with common initial and terminal strings in English, Russian, and Italian (Tables 4.14, 4.15, 6.6, 6.11). These resources will be of use to pharmaceutical manufacturers proposing new names and to the INN Expert Committee responsible for evaluating and approving them. In summary, these contributions significantly enhance the knowledge base of terminology, drug nomenclature, LASA errors, and pharmacovigilance.

7.4.2 Implications for practice

This chapter discusses the findings of the thesis in terms of answering the three research questions. In discussing the findings from the statistical analysis (Research Question 2), findings were considered in light of relevant suggestions and considerations for manufacturers proposing INNs and the WHO INN Expert Committee approving those INNs. Findings can contribute to the naming process by recommending resources and quantitative measures to assess a proposed name's risk of LASA error and pinpointing linguistic properties to be avoided or encouraged. Implications for practice are summarised in this section.

It is important to note that these implications are broadly based on pilot evidence that suggested an association between certain linguistic features and the risk of error, but they are not supported by sufficient empirical evidence to make firm recommendations for policymakers (they are framed as *considerations*). One finding, however, is supported by enough empirical evidence to suggest a change in policy (to increase diversity in the beginning strings of INNs), but in practice this feature is steadily improving over time so the risk is reduced with newly approved names (this is framed as a *suggestion*).

Suggestion for policy: encourage diversity of strings used in the beginning of names. Encouragingly, the beginnings of INNs are becoming more diverse over time, and this diversity should be promoted in the INN name approval process. Future names would benefit from avoiding the use of common beginnings such as pro-. This is more feasible when the string is not part of the STEM, which in most cases is at the end of the name. The list of frequent initial strings in Table 4.14 is hopefully a useful resource for the INN Expert Committee when devising new PREFIX STEMS.

Consideration for policy: review the list of prohibited letters, in particular <k>. WHO guidelines prohibit the use of certain letters or groups of letters, to facilitate translation and international understanding of INNs. The case studies of Russian transl(iter)ation and Italian translation in Chapter 6 found that not all prohibited letters pose challenges, and some in fact serve to disambiguate homographs in English that represent different phonemes, e.g. hard /k/ and soft /s/, which can both be represented with <c> in English (see acamprosate and amoxicillin). Indeed, the letter <k> is in several SUBSTEMS in recently-devised naming schemes for biologics, e.g. etakafusp alfa, so it isn't going anywhere and the ban on its usage should be reviewed.

Consideration for policy: discourage the use of the ending -ine in new pharmacological stems. As with beginnings of INNs, their endings are becoming more diverse over time, and this diversity should be promoted in the INN name approval process. In particular, INNs are dominated by the ending -ine, in a fifth of the names analysed here and accounting for more than 30% of names involved in LASA errors. It is challenging to reduce the use of certain strings at the endings of names, because the majority of STEMS are SUFFIX STEMS. However, the use of -ine is falling over time and is it promising that no newly created or proposed STEMS in the most recent WHO 'Pre-stems' publication end in -ine (WHO, 2025), although of course existing names persist. The list of frequent initial strings in Table 4.15 is hopefully a useful resource for the INN Expert Committee when devising new SUFFIX STEMS. The finding on an association between terminal trigrams and error is not as robust as the finding on initial trigrams. However, simply discouraging the use of -ine in new STEMS would halt the dominance of this ending in the nomenclature and promote diversity in endings.

Consideration for policy: note the STEM positioning of names in a pharmacological group. Findings here suggest that semantic transparency of names is steadily increasing over time. Although the association of this linguistic feature with error was not certain, the positive trend of increasing transparency in names' pharmacological group at the very least benefits users' recognition and recall of names, and has the potential to protect against error. When manufacturers are proposing new INNs it is suggested that they consider where the STEM is positioned in the proposed name and if this is consistent with the position of STEMS in the other names in that pharmacological group.

Anecdotally, prescribers report that they do not decompose drug names into their STEMS and SUBSTEMS, though they likely develop some intuitive understanding of STEMS over time. A lack of understanding leads to workarounds to decipher names (Rupprecht and Rhodes 2014), but users of drug names would benefit from some formal training in drug nomenclature. The WHO has created the School of INN, an autodidactic course in INN naming, which goes some way to educate users in the designation process (SoINN, 2020b). Prescribers would benefit from a searchable database of names that highlight STEMS and the underlying taxonomy. Unfamiliar names have been found to be misperceived more than familiar names (Lambert et al., 2010), and this could be addressed by increasing users' knowledge of the nomenclature.

To conclude, it is better to optimise a name before its approval rather than release a name into the system with high-risk features and have to mitigate against its involvement in LASA errors. INNs are rarely changed, with only two documented cases of name changes. One occurred over 20 years ago and lacked documentation, while a recent change involved a name that sounded offensive in Swedish and was altered a year post-approval (upifitamab to upinitatug). The change took place around a year after approval but before the drug's market authorisation. Addressing problematic names post-approval becomes the responsibility of pharmacovigilance agencies, including those managing LASA error watch lists, such as the Institute for Safe Medication Practices (ISMP).

The finding in this thesis that prescribing frequency is a significant risk factor for LASA error is meaningful for those managing lists of LASA name pairs. Any linguistic latent conditions in a name would be outstripped by a high prescribing frequency - a high-frequency name with a low probability of error could result in more errors than a low-frequency name with a high probability of error (Lambert, Chang, & Gupta, 2003). Similarly, names prescribed early in the programme have an increased risk of LASA error, because they have been in circulation for longer and this has created opportunities for error. The problem is that older names were not subject to the current rules when they were coined and yet they persist in the nomenclature. It may be that rules have been variably applied over the years, explaining a lot of the inconsistencies. Names prescribed in the first two decades of the programme should be considered for inclusion in LASA name pair lists.

7.4.3 Implications for research and future research directions

The case studies in Chapter 8 explore transliteration and translation of English-language INNs into Russian-language INNs, and translation of English-language INNs into Italian Common names. These case studies provide an analytical framework for researchers of nomenclatures in other languages.

The predominance of prescribing frequency and year of approval should be taken into account in any studies of the causative factors of LASA error. The finding runs counter to current wisdom that prescribing frequency is associated with improved recall of drug names (Lambert et al. 2003), purely because of the increased number of opportunities for an often-prescribed drug to be involved in an error.

Findings from the statistical analysis lend support for certain linguistic features being associated with error, but fall short of the evidence burden required to make policy

recommendations. Therefore, they merit further research endeavours to explore their association with LASA error, or indeed to explore linguistic features using new methods or in other nomenclatures. These are discussed in detail in section 7.2 earlier in this chapter, and again briefly here. Future research could disaggregate the prohibited letters (Orth-03) used in names and explore their association with error; this would quantitatively support any review of prohibited letters taken by the INN Expert Committee. Studies could explore the relationship between various characteristics of STEM taxa (Morph-09) and LASA error, as well as the relationship between semantic transparency and LASA error (Rel-10), potentially analysing other drug nomenclatures for comparison or by adapting the criteria of each type to allow deeper understanding of STEM taxon features. Future research could look at the impact on error prediction of weighting the name at the end (Rel-04) as well as at the beginning. Furthermore, the pilot evidence merits further exploration of neighbourhood frequency and its relationship with LASA error – this could include assessing ways to measure neighbourhood frequency and a feasibility study into incorporating this measure into the INN Expert Committee's decision-making process.

More broadly, there are several promising directions for future research of this topic. First, drug nomenclatures in other languages should be characterised, to provide insights into how names are translated into other writing systems or language branches. Since the INN form is used as default in most national systems, it would be most fruitful to begin this process by exploring the INN nomenclatures in the other UN working languages: Arabic, Chinese, French, and Spanish. Standardised translation guides and tools for translating and transliterating drug names into various languages could be developed. We need data on LASA errors in languages other than English in order to assess the association between linguistic features and error, and more broadly to develop LASA error name watch lists for other nations. Exploring phonological similarity and stress patterns in different languages, as well as how stress and pronunciation vary based on the speaker's language background, could offer insights into LASA errors caused by phonological similarity.

Further investigation of the alignment between the ATC index and the INN taxonomy is needed to understand whether alignment of these two systems is of tangible importance in patient safety. The analysis could be expanded to UK combination names, e.g. Co-amoxiclav (amoxicillin, a penicillin-type antibiotic, and clavulanic acid, a beta-lactamase inhibitor). Granular error data including frequency, type, clinical setting, and language could be attained using data mining techniques. A framework could be developed to assess the three axes of error presented in Chapter 3: opportunity, probability, and intensity. The framework would include seriousness of outcome and addresses a lack of consistent

terminology in the literature to describe the functions of an error (Ferner and Aronson 2006; Aronson and Ferner 2005). Future work in this field could update the analysis to include recently-approved biologics or proposed names, leading to the potential development of algorithms for testing proposed names.

Names that are neither right- nor left-headed make up almost a third of the database. These names place the STEM in the middle of the name, followed by a SUBSTEM (agglutinated), or the SUBSTEM replaces the STEM (replacement). This phenomenon was also observed in the Russian and Italian nomenclatures, because most STEMS do not change their position in translation. Although the association of agglutination vs replacement with LASA errors was not measured, it merits further study, considering the association found here between STEM position, taxon shape, and LASA errors.

Other future work could examine pairwise linguistic features of INNs and their association with LASA error. In this study standalone measurements were taken but pairwise measurements are more accurate because they take both names involved in the error into account and can identify salient asymmetric features. This pairwise data analysis could consider confusable trios with asymmetric similarities, such as carboplatin, cisplatin and carboprost, first highlighted in Bryan et al., (2020b). Lastly, future research could include conducting a sensitivity analysis to test the robustness of the findings in Chapter 5 on risk factors to variations in assumptions about how these factors influence error risk.

7.4.4 Limitations

This thesis has several limitations related to data availability, methodological constraints, and analytical challenges, all of which have implications for the generalisability and interpretation of findings.

One key limitation is the scope of the study, which was shaped by both data availability and the need for a focussed analysis. The thesis examines generic medication names (not brand names) in the English-language INN nomenclature published by the WHO, as well as Russian-language INNs and Italian Common Names (DCIts). However, multi-word names were excluded to allow for comparability, following the approach in Bryan et al. (2015). While this decision helps control for factors like name length, it runs the risk of overlooking complexities introduced by modifiers such as *sodium* or *acid*. Additionally, the dataset for statistical analysis was restricted to INNs authorised in the UK, excluding over 80% of the full

INN nomenclature. This was necessary because unauthorised names have zero probability of being involved in clinical LASA errors, but it may introduce selection biases.

The research questions were designed to be sequential, with RQ1 (descriptive linguistic analysis) informing RQ2 (statistical analysis of linguistic properties and LASA error risks), and both RQ1 and RQ2 informing RQ3 (comparative analysis with Russian and Italian names). However, RQ1 is limited to English because the WHO Stem Book (which provides guidance on pharmacological STEMS) is only published in English. Similarly, RQ2 is also restricted to English due to the availability of LASA error reports solely in this language. RQ3 explores Russian-language INNs and Italian generic names but does not include LASA errors or morphosemantic features, given the lack of comparable data.

A further limitation relates to linguistic and subjective constraints. The thesis is shaped by the language proficiency of the researcher (RB), who is a native English speaker, fluent in Italian, with basic proficiency in Russian, Spanish, and Welsh. This influenced the analysis of pronunciation and syllable counts, which were inherently subjective and could differ from those of speakers of other English dialects or non-native English speakers. Additionally, the INN nomenclature has evolved over seven decades (1955–2019), incorporating diachronic changes in naming conventions and prescribing patterns. These changes add complexity and potential biases, as naming trends from different periods are compared within a single analytical framework.

The statistical analysis also has several key limitations. One constraint is the binary nature of the outcome measure (LASA error), which was dictated by data availability. This prevents a more nuanced understanding of LASA error drivers, as the analysis does not distinguish between names that have been confused frequently and those that have been confused only a handful of times. More granular data – such as error frequency, severity, clinical setting, consequences, and the first language of the healthcare professional – would have provided deeper insights into the factors influencing error risk. Additionally, the study does not account for non-linguistic variables such as stress levels, urgency, and contextual factors, all of which could significantly impact LASA errors.

Another challenge arises from confounding variables. The statistical model is heavily influenced by two non-linguistic factors: prescribing frequency and year of approval. Because these variables dominate the risk of LASA errors, even small modelling imperfections could have significant consequences. Furthermore, collinearity between some variables makes it difficult to separate direct and indirect effects, complicating the

interpretation of results. Non-random missing data further adds to the complexity, as it introduces the possibility of bias.

The statistical approach also presents methodological challenges. Stepwise selection with backward elimination was used, which carries the risk of overfitting the training data and reducing the model's generalisability. This method does not necessarily include all relevant factors and may not fully translate to real-world clinical scenarios. However, the final model appears robust to resampling which gives greater weight to the findings. Additionally, issues such as high collinearity in the global model and a high 'events per variable' (EPV) ratio required countermeasures such as bootstrapping, yet some level of statistical uncertainty remains.

Finally, there are limitations related to the definition of a drug name's neighbourhood. Neighbourhood frequency was not restricted to a single clinical specialty, meaning that names for drugs used in entirely different therapeutic areas were included. While this approach ensured consistency in the modelling, it may have led to the inclusion of names that, in practice, are unlikely to be confused.

Despite these limitations, the study makes a significant contribution by addressing gaps in the literature and offering new insights into the linguistic and statistical properties of INN nomenclature. The constraints outlined above highlight areas for future research, particularly in incorporating additional contextual factors, expanding linguistic comparisons, and refining statistical modelling approaches.

7.4.5 Summary and closing remarks

In this study I have analysed linguistic features of the English-language INN nomenclature and their association with LASA error, identifying one **linguistic latent condition** that primes the risk of LASA errors (frequency of the name's initial trigram), and six other **potential linguistic latent conditions** that warrant further research and exploration. I have also identified two non-linguistic confounders that significantly drive risk: prescribing frequency and year of approval. Descriptive and statistical analyses are complemented by two case studies of Russian-language INNs and Italian Common Names, including a consideration of implications of findings to these nomenclatures.

Core findings of this study speak to the importance of diversity within the INN nomenclature and uniformity within STEM taxa of the nomenclature. Diversity in the strings used in the

beginnings – and to some extent the endings – of names reduces the pairwise orthographic similarity between names, thus reducing the number of confusable pairs. Complexity and diversity in the ways in which STEM taxa are structured or STEMS are positioned in names could possibly reduce error by prompting users to pay attention, but this needs further study. Findings also point to uniformity *within* a STEM taxon being important to the reduction of error, i.e. that STEMS should be in a predictable and consistent position for their STEM taxon.

The overarching goal of this study is to enhance patient safety. In making the drug lexicon safer, errors are mitigated and the number or seriousness of harmful outcomes can be reduced. Although many of the historical advancements in pharmacovigilance have been prompted by 'misfortune or disaster' (see the case of thalidomide in McBride 1961; other examples in Goldberg 1987; Routledge 1998), it is surely preferable to put adequate safeguards in place ahead of time to prevent error proactively. This approach aligns with the stance of the World Health Organization (WHO) and conventional wisdom in pharmacovigilance, that medication errors are inevitable and humans are fallible, and that it is the responsibility of regulators to implement sufficient safeguards to mitigate risks.

This thesis is embedded in a linguistic problem that can cause harm to patients and even result in death (Tuohy and Paparella, 2005, Bryan et al., 2020a), and I have used novel analytical techniques to generate new knowledge on drug names and LASA errors and ultimately to recommend changes to policy on drug naming. INNs are international names, so I felt it was important to explore naming and contextualise my findings in languages other than English. Among many interesting findings, I was surprised to learn that despite right-headed compounds being the predominant form in English and in neoclassicals, right-headedness is not protective of error and other forms of headedness such as left-headedness and exocentricity (not displaying the semantic head) might in fact be beneficial. In identifying linguistic drivers of error as latent conditions in the healthcare system I have highlighted an opportunity to change policy on drug naming that can improve patient safety. I hope to work with the INN Expert committee to address these recommendations and put them into practice. I'm excited to work with them in the future to further enhance the safety of patients.

Appendix[50]

A1 Glossary of terms

Term	Definition	Source
Medicinal product	(a) any substance or combination of substances presented as having properties for treating or preventing disease in human beings; or (b) any substance or combination of substances which may be used in or administered to human beings either with a view to restoring, correcting or modifying physiological functions by exerting a pharmacological, immunological or metabolic action or to making a medical diagnosis.	EU directive 2001/83/EC (EU 2001:13)
Medication and medicinal product	The treatment process; the act of prescribing (deciding what they prescribe), writing a prescription, dispensing, transcribing, or administering a medicinal product.	Medication, medicine, drug, and medicinal product are often used interchangeably. Medication is particularly ambiguous as it is both the process, and the object undergoing the process. In this review we opt to use *medicinal product* to refer to the object, and medication to refer to the process.
Medicine	The practice and theory of medicine	Common usage in MeSH
Drug	Common attributive usage, as in “drug-induced” and commonly collocated with adverse, such as in “adverse drug reaction”.	Drug is commonly collocated as “wrong drug”, in the ‘5 Rs’ framework, and this is maintained here, such as in ‘wrong drug error’. Common usage in MeSH.
Error	Something incorrectly done through ignorance or inadvertence; a mistake, e.g. in calculation, judgement, speech, writing, action, etc. or a failure to complete a planned action as intended, or the use of an incorrect plan of action to achieve a given aim.	Cited from Aronson 2009a
Medication error	A failure in the drug treatment process that leads to, or has the potential to lead to, harm to the patient.	Where ‘medication’ is defined as above (adapted from Aronson 2009a, to add reference to the use of a drug)

[50] Appendices comprise just over 20,000 words – thesis chapters 1 through 7 comprise just under 90,000 words.

A2 Searches undertaken for narrative review of literature in health science, and case reports

English language search, run 14.6.19, 1,280 results

KEY: [ti] = Title field only; [tiab] = Title and Abstract fields; MESH = Medical Subject Headings (MESH). No Limits Applied.
#1 prescri*[tiab] OR drug administ*[tiab] OR dispens*[tiab] OR pharmacist*[tiab] OR drug*[tiab] OR medicine*[tiab] OR medication*[tiab] OR pharmaceutical*[tiab] OR pharmacotherapeutic agent*[tiab] OR medicinal product*[tiab] OR "Drug Therapy"[Mesh] OR "Prescriptions"[Mesh] OR "Pharmaceutical Preparations"[Mesh]
#2 error*[tiab] OR mix* up[tiab] OR "mix ups"[tiab] near* miss*[tiab] OR oversight*[tiab] OR wrong*[tiab] OR mistake*[tiab] OR suboptimal practice[tiab] OR “Medical Errors”[mh]
#3 (#1 AND #2)
#4 “Medication Errors”[mh]
#5 drug incident*[tiab] OR medication incident*[tiab] OR medication error*[tiab] OR drug* error*[tiab]
#6 (#3 OR #4 OR #5)
#7 (lookalike[tiab] OR “look-alike*”[tiab] OR soundalike[tiab] OR sound-alike*[tiab] OR LASA[tiab] OR LASAs[tiab]
#8 similarity[tiab] OR similar[ti]
#9 confus*[tiab]
#10 ((name*[tiab] OR title*[tiab]) AND (generic[tiab] OR brand*[tiab] OR trade[tiab] OR proprietary[tiab] OR commercial[tiab] OR international nonproprietary[tiab] OR INN[tiab] OR INNs[tiab] OR BAN[tiab] OR BANs[tiab] OR British approved[tiab] OR pharmacopoeial title*[tiab] OR nomenclature*[tiab] OR package*[tiab] OR label*[tiab]))
#11 (#7 OR #8 OR #9 OR #10)
#12 (#11 AND #6)
#13 animals[mh] NOT humans[mh]
#14 (#12 NOT #13)

Russia-specific Pubmed search, run 13.8.19, zero results

KEY: [ti] = Title field only; [tiab] = Title and Abstract fields; MESH Medical Subject Headings (MESH). No Limits Applied.
#1 prescri*[tiab] OR drug administ*[tiab] OR dispens*[tiab] OR pharmacist*[tiab] OR drug*[tiab] OR medicine*[tiab] OR medication*[tiab] OR pharmaceutical*[tiab] OR pharmacotherapeutic agent*[tiab]

OR medicinal product*[tiab] OR "Drug Therapy"[Mesh] OR "Prescriptions"[Mesh] OR "Pharmaceutical Preparations"[Mesh]
#2 error*[tiab] OR mix* up[tiab] OR "mix ups"[tiab] near* miss*[tiab] OR oversight*[tiab] OR wrong*[tiab] OR mistake*[tiab] OR suboptimal practice[tiab] OR “Medical Errors”[mh]
#3 (#1 AND #2)
#4 “Medication Errors”[mh]
#5 drug incident*[tiab] OR medication incident*[tiab] OR medication error*[tiab] OR drug* error*[tiab]
#6 (#3 OR #4 OR #5)
#7 (lookalike[tiab] OR “look-alike*”[tiab] OR soundalike[tiab] OR sound-alike*[tiab] OR LASA[tiab] OR LASAs[tiab]
#8 similarity[tiab] OR similar[ti]
#9 confus*[tiab]
#10 (name*[tiab] OR title*[tiab]) AND (generic[tiab] OR brand*[tiab] OR trade[tiab] OR proprietary[tiab] OR commercial[tiab] OR international nonproprietary[tiab] OR INN[tiab] OR INNs[tiab] OR BAN[tiab] OR BANs[tiab] OR British approved[tiab] OR pharmacopoeial title*[tiab] OR nomenclature*[tiab] OR package*[tiab] OR label*[tiab])
#11 (#7 OR #8 OR #9 OR #10)
#12 (Russia*[tiab] OR Росси*[tiab])
#13 (#11 AND #6 AND #12)
#14 animals[mh] NOT humans[mh]
#15 (#13 NOT #14)

Italy-specific Pubmed search, run 13.8.19, 6 results

KEY: [ti] = Title field only; [tiab] = Title and Abstract fields; MESH = Medical Subject Headings (MESH). No Limits Applied.
#1 prescri*[tiab] OR drug administ*[tiab] OR dispens*[tiab] OR pharmacist*[tiab] OR drug*[tiab] OR medicine*[tiab] OR medication*[tiab] OR pharmaceutical*[tiab] OR pharmacotherapeutic agent*[tiab] OR medicinal product*[tiab] OR "Drug Therapy"[Mesh] OR "Prescriptions"[Mesh] OR "Pharmaceutical Preparations"[Mesh]
#2 error*[tiab] OR mix* up[tiab] OR "mix ups"[tiab] near* miss*[tiab] OR oversight*[tiab] OR wrong*[tiab] OR mistake*[tiab] OR suboptimal practice[tiab] OR “Medical Errors”[mh]
#3 (#1 AND #2)
#4 “Medication Errors”[mh]
#5 drug incident*[tiab] OR medication incident*[tiab] OR medication error*[tiab] OR drug* error*[tiab]
#6 (#3 OR #4 OR #5)
#7 (lookalike[tiab] OR “look-alike*”[tiab] OR soundalike[tiab] OR sound-alike*[tiab] OR LASA[tiab] OR LASAs[tiab]

#8 similarity[tiab] OR similar[ti]
#9 confus*[tiab]
#10 (name*[tiab] OR title*[tiab]) AND (generic[tiab] OR brand*[tiab] OR trade[tiab] OR proprietary[tiab] OR commercial[tiab] OR international nonproprietary[tiab] OR INN[tiab] OR INNs[tiab] OR BAN[tiab] OR BANs[tiab] OR British approved[tiab] OR pharmacopoeial title*[tiab] OR nomenclature*[tiab] OR packag*[tiab] OR label*[tiab])
#11 (#7 OR #8 OR #9 OR #10)
#12 (Itali*[tiab])
#13 (#11 AND #6 AND #12)
#14 animals[mh] NOT humans[mh]
#15 (#13 NOT #14)

Italian-language Pubmed search, run 30.09.19, 0 results

KEY: [ti] = Title field only; .[tiab] = Title and Abstract fields; MESH = Medical Subject Headings (MESH). No Limits Applied.
#1 prescri*[tiab] OR ricett*[tiab] OR somministrazion*[tiab] OR dispens*[tiab] OR farmacist*[tiab] OR medicin*[tiab] OR farmac*[tiab] OR medicinal*[tiab] OR farmaceutic*[tiab] OR "Drug Therapy"[Mesh] OR "Prescriptions"[Mesh] OR "Pharmaceutical Preparations"[Mesh]
#2 error*[tiab] OR scambi* [tiab] OR disguid* [tiab] OR equivoc* [tiab] OR sbagli*[tiab] OR wrong*[tiab] OR disattenzion*[tiab] OR quasi-error*[tiab] OR manca* poco[tiab] OR “Medical Errors”[mh]
#3 (#1 AND #2)
#4 “Medication Errors”[mh]
#5 Incident* medicina*[tiab] OR incident* farmac*[tiab] OR farmac* sbagliat*[tiab] OR error* medicina*[tiab] OR error* farmac*[tiab]
#6 (#3 OR #4 OR #5)
#7 (lookalike[tiab] OR “look-alike*”[tiab] OR soundalike[tiab] OR sound-alike*[tiab] OR LASA*[tiab] OR sembra* simil*[tiab] OR suona* simil*[tiab] OR grafic*[tiab] OR fonetic*[tiab]
#8 somiglianz*[tiab] OR simil*[ti]
#9 confond*[tiab] OR confus*[tiab]
#10 (nom*[tiab] OR titol*[tiab]) AND (generic[tiab] OR brand*[tiab] OR marchi[tiab] OR commercial[tiab] OR denominazion* comun* italian*[tiab] OR DCIt[tiab] OR denominazion* comun* internazional*[tiab] OR terminolog*[tiab] OR nomenclatur*[tiab]
#11 (#7 OR #8 OR #9 OR #10)
#12 (#11 AND #6)
#13 animals[mh] NOT humans[mh]
#14 (#12 NOT #13)
#15 Search Italian[Language]

#16 (#14 AND #15)

Russian-language Pubmed search, run 30.09.19, 0 results

KEY: [ti] = Title field only; [tiab] = Title and Abstract fields; MESH = Medical Subject Headings (MESH). No Limits Applied.
#1 Приготовл*[tiab] OR отпуск*[tiab] OR фармац*[tiab] OR лекарств*[tiab] OR медикамент*[tiab] OR проход*[tiab] OR "Drug Therapy"[Mesh] OR "Prescriptions"[Mesh] OR "Pharmaceutical Preparations"[Mesh]
#2 ошиб*[tiab] OR путаниц*[tiab] OR недосмотр*[tiab] OR неправильн*[tiab] OR ошиб*[tiab] OR “Medical Errors”[mh]
#3 (#1 AND #2)
#4 “Medication Errors”[mh]
#5 инцидент* лекарств*[tiab] OR ошиб* лекарств*[tiab]
#6 (#3 OR #4 OR #5)
#7 (двойн*[tiab] OR LASA*[tiab]
#8 сходств*[tiab] OR подобн*[ti] OR похож*[tiab]
#9 пута*[tiab]
#10 (им*[tiab] OR названи*[tiab]) AND (общ*[tiab] OR Фирменн* Марк*[tiab] OR коммерческ*[tiab] OR Международн* непатентованн* наименован*[tiab] OR INN*[tiab] OR номенклатур*[tiab])
#11 (#7 OR #8 OR #9 OR #10)
#12 (#11 AND #6)
#13 animals[mh] NOT humans[mh]
#14 (#12 NOT #13)
#15 Search russian[Language]
#16 (#14 AND #15)

A3 Searches undertaken for systematic review (Bryan et al., 2020a)

Terms were searched in title and abstract ([tiab] on PubMed) to retrieve results that are both focussed (title) and closely related (abstract) to LASA errors. First several broad parameters that may co-occur with LASA error discourse, such as the method of delivery (e.g. dispensing), the type of incident (e.g. near miss), medication or a synonym (e.g. drug), the type of medication name (e.g. generic), aspects such as packaging or labelling, and interventions such as font adaptation and automated alerts (e.g. Tall Man lettering) were defined. These were then combined with Boolean operator AND to form an array of search terms. This array was combined with variants of LASA to find more specific papers, and the

search was finally narrowed by searching for only those papers fulfilling the preceding conditions and with relevant MeSH terms as a major subject heading([majr] on PubMed).

PubMed search strategy, run 16.12.15, 4,168 results

KEY: [ti] = Title field only; [tiab] = Title and Abstract fields; MESH = Medical Subject Headings. No limits applied.

#1 prescri*[tiab] OR drug administ*[tiab] OR dispens*[tiab] OR pharmacist*[tiab] OR drug*[tiab] OR medicine*[tiab] OR medication*[tiab] OR pharmaceutical*[tiab] OR pharmacotherapeutic agent*[tiab] OR medicinal product*[tiab] OR "Drug Therapy"[Mesh] OR "Prescriptions"[Mesh] OR "Pharmaceutical Preparations"[Mesh]

#2 Error*[tiab] OR mix* up[tiab] OR "mix ups"[tiab] near* miss*[tiab] OR oversight*[tiab] OR wrong*[tiab] OR mistake*[tiab] OR suboptimal practice[tiab] OR “Medical Errors”[mh]

#3 (#1 AND #2)

#4 “Medication Errors”[mh]

#5 drug incident*[tiab] OR medication incident*[tiab] OR medication error*[tiab] OR drug* error*[tiab]

#6 (#3 OR #4 OR #5)

#7 (lookalike[tiab] OR “look-alike*”[tiab] OR soundalike[tiab] OR sound-alike*[tiab] OR LASA[tiab] OR LASAs[tiab]

#8 similarity[tiab] OR similar[ti]

#9 confus*[tiab]

#10 ((name*[tiab] OR title*[tiab]) AND (generic[tiab] OR brand*[tiab] OR trade[tiab] OR proprietary[tiab] OR commercial[tiab] OR international nonproprietary[tiab] OR INN[tiab] OR INNs[tiab] OR BAN[tiab] OR BANs[tiab] OR British approved[tiab] OR pharmacopoeial title*[tiab] OR nomenclature*[tiab] OR packag*[tiab] OR label*[tiab]))

#11 (colour[tiab] OR color[tiab] OR font[tiab] OR fonts[tiab] OR typograph*[tiab] OR Tall Man[tiab] OR tall man[tiab] OR lettering[tiab] OR upper case[tiab] OR lower case[tiab] OR automated alert*[tiab] OR CPOE*[tiab] OR computerized physician order entry[tiab] OR barcod*[tiab] OR distract*[tiab] OR interrupt*[tiab])

#12 (#7 OR #8 OR #9 OR #10 OR #11)

#13 (#12 AND #6)

#14 animals[mh] NOT humans[mh]

#15 (#13 NOT #14)

Web of Science search strategy, run 26.01.16, 2,126 results

KEY: TS= Topic Search (Title, Abstract, and Keywords); TI= Title Only Search. No limits applied.
#1 TS=(prescri* OR “drug administ*” OR dispens* OR pharmacist* OR drug* OR medicine* OR medication* OR pharmaceutical* OR “pharmacotherapeutic agent*” OR “medicinal product*”)
#2 TS=(error* OR “mix* up” OR “mix-up*” OR “near* miss*” OR oversight* OR wrong* OR mistake* OR “suboptimal practice” OR “medical error*”)
#3 (#1 AND #2)
#4 TS=(“drug incident*” OR “medication incident*” OR “medication error*” OR “drug error*”)
#5 (#4 OR #3)
#6 TS= (lookalike* OR “look-alike*” OR soundalike* OR “sound alike*” OR LASA OR LASAs OR "LA-SA" OR "LA-SAs")
#7 TI= (similar)
#8 TS=(similarity OR similarities OR confus*)
#9 (#6 OR #7 OR #8)
#10 TS=(name* OR title*)
#11 TS=(generic OR brand* OR trade OR proprietary OR commercial)
#12 TS=(“international nonproprietary” OR “international non-proprietary” OR INN OR INNs OR BAN OR BANs OR “British approved” OR pharmacopoeial OR nomenclature OR Packag* OR label*)
#13 (#10 AND (#11 OR #12)
#14 (TS=(colo*r OR font OR fonts OR typograph* OR “tall man” OR Tall Man OR lettering OR “lower case” OR “upper case” OR capitali?ation OR “automated alert” OR CPOE OR “computeri?ed physician order entry” OR barcod* OR distract* OR interrupt*)
#15 (#9 OR #13 OR #14)
#16 (#15 AND #5)
#17 (Exclude: Letters, Editorials, News Items, Notes, Item about an Individual, Book Chapter)

EMBASE search strategy (via OVID), run 16.12.15, 3,382 results

KEY: .ti. = Title field only; ti,ab. = Title and Abstract fields; / = EMTREE index term. No limits applied.
#1 (prescri* or drug administ* or dispens* or pharmacist* or drug* or medicine* or medication*).ti,ab.
#2 (pharmaceutical* or pharmacotherapeutic agent* or medicinal product*).ti,ab.
#3 drug administration/
#4 prescription
#5 drug/
#6 drug therapy/
#7 (#1 or #2 or #3 or #4 or #5 or #6)
#8 (error* or mix* up or mix-up* or near* miss* or oversight* or wrong* or mistake* or suboptimal practice).ti,ab.
#9 medical error/ or error/

#10 (#8 or #9)
#11 (#7 and #10)
#12 medication error/
#13 (drug incident* or medication incident* or medication error* or drug error*).ti,ab.
#14 (#11 or #12 or #13)
#15 (lookalike* or look-alike* or soundalike* or sound alike* or LASA or LASAs or LA-SA or LA-SAs).ti,ab.
#16 similar.ti.
#17 (similarity or similarities or confus*).ti,ab.
#18 (#16 or #17)
#19 (name* or title*).ti,ab.
#20 (generic or brand* or trade or proprietary or commercial).ti,ab. (similarity or similarities or confus*).ti,ab.
#21 generic drug
#22 (international nonproprietary or international non-proprietary or INN or INNs or BAN or BANs or British approved or pharmacopoeial or nomenclature or Packag* or label*).ti,ab.
#23 *drug labeling/
#24 *drug packaging/ or *packaging/
#25 (#20 or #21 or #22 or #23 or #24)

Scopus search strategy, run 04.02.16, 2,344 results

KEY: TITLE = Title field only; TITLE-ABS-KEY = Title, abstract and keyword fields (Health Sciences and Social Sciences and Humanities only). No other Limits applied except for final exclusion as per #17.
#1 TITLE-ABS-KEY(prescri* OR “drug administ*” OR dispens* OR pharmacist* OR drug* OR medicine* OR medication* OR pharmaceutical* OR “pharmacotherapeutic agent*” OR “medicinal product*”)
#2 TITLE-ABS-KEY(error* OR “mix* up” OR “mix-up*” OR “near* miss*” OR oversight* OR wrong* OR mistake* OR “suboptimal practice” OR “medical error*”)
#3 #1 AND #2
#4 TITLE-ABS-KEY(drug incident*” OR “medication incident*” OR “medication error*” OR “drug error*”)
#5 #4 OR #3
#6 TITLE-ABS-KEY(lookalike* OR “look-alike*” OR soundalike* OR “sound alike*” OR LASA OR LASAs OR "LA-SA" OR "LA-SAs")
#7 TITLE(similar)
#8 TITLE-ABS-KEY(similarity OR similarities OR confus*)

#9 #6 OR #7 OR #8

#10 TITLE-ABS-KEY(name* OR title*)

#11 TITLE-ABS-KEY(generic OR brand* OR trade OR proprietary OR commercial)

#12 TITLE-ABS-KEY(“international nonproprietary” OR “international non-proprietary” OR INN OR INNs OR BAN OR BANs OR “British approved” OR pharmacopoeial OR nomenclature OR Packag* OR label*)

#13 #10 AND (#11 OR #12)

#14 TITLE-ABS-KEY(colo*r OR font OR fonts OR typograph* OR “tall man” OR Tall Man OR lettering OR “lower case” OR “upper case” OR capitali?ation OR “automated alert” OR CPOE OR “computeri?ed physician order entry” OR barcod* OR distract* OR interrupt*)

#15 #9 OR #13 OR #14

#16 #15 AND #5

#17 Exclude: Letters, Editorials, Notes, Book Chapter, Books

A4 Summary of medication error prevalence

		Prescribing	Transition	Dispensing	Administration	Monitoring	TOTALS
Setting							
Primary care errors	n	43,623,500		32,877,883		14,503,519	91,004,902
	%	47.90%		36.10%		15.90%	100.00%
Secondary care errors	n	4,043,745	3,380,383	1,376,609	37,258,284	1,368,644	47,427,665
	%	8.50%	7.10%	2.90%	78.60%	2.90%	100%
Care home errors	n	2,983,219		3,522,355	91,832,832	625,398	98,963,804
	%	3.00%		3.60%	92.80%	0.60%	100.00%
Seriousness							
Little or no potential for clinical harm	n	24,281,668	1,390,365	24,469,095	119,282,765	1,799,734	171,223,627
	%	48.00%	41.10%	64.80%	92.40%	10.90%	72%
Potential to cause moderate harm	n	25,296,213	1,744,801	12,878,470	9,447,407	11,998,226	61,365,117
	%	49.90%	51.60%	34.10%		72.70%	26%
Potential to cause severe harm	n	1,072,583	245,217	429,282	360,944	2,699,601	4,807,627
	%	2.10%	7.30%	1.10%		16.40%	2%
TOTAL ERRORS	n	50,650,464	3,380,383	37,776,847	129,091,116	16,497,561	237,396,371
	%	21.30%	1.40%	15.90%	54.40%	6.90%	100.00%

A5 Communications log

5.10.15	Emailed Clare Boucher, subject librarian at SU requesting access to the European Pharmacopoeia
26.10.15	Email helpdesk@aifa.gov.it requesting access to a DCIt database, either as the EP or in some other format
20.5.16	Emailed herve.le-louet@hmn.aphp.fr (I met him in the conference in Sicily) regarding resources the EMA could offer access to
23.5.16	Sue J emailed Amanda Julie Neville (nvm@unife.it) asking for access to DCIt database, perhaps in the Italian Pharmacopoeia
23.5.16	Amanda replied recommending the informatore farmaceutico – this is basically medicines.org.uk
26.5.16	RB found the farmaci.agengiafarmaoc.gov.it. This is officially monitored by the EMA and the Italian Pharmacy Agency.
1.6.16	I emailed INN secretariat Raffaella Giovanna Balocco requesting access to parallel nomenclatures – she emailed back to decline, but did pass on the recent paper to the INN Expert Group.
4.6.16	I emailed INN secretariat asking for more information on how STEMS are parsed in other languages. No reply.
1.12.2016-31.3.2019	Extended maternity leave
10.7.19	Requested access to the ISMP database
17.7.19	Farzana Samad from the ISMP replied – I cannot have access as it is not a public database. She also cautioned that as the ISMP is not a regulatory body, reporting is optional and thus it won't give an accurate representation of error frequency. She recommended making a user account to read case reports in more detail.
2.8.19	Discussed with Jeffrey Aronson getting access to report data from the MHRA or EMA.
June 2021	Discussion translation and transliteration with INN Secretariat
May 2023	Discussed community and hospital prescribing data with NHS Digital

A6 List of confused error pairs used as outcome measure in statistical analysis

Intended name	Confused name	Reciprocal pair?	ISMP, 2019	MER, 2000	Galanter et al., 2014	Basco et al., 2016	Lambert et al., 2019	McNulty and Spurr, 1979	Chisholm & Klanduch 2000	Heck et al. 2020	Cohen & Smetzer 2014
acetazolamide	acetylcysteine	no				*					
allopurinol	atenolol	no				*					
alprazolam	lorazepam	yes	*	*							
alprazolam	lansoprazole	no			*						
amiloride	amlodipine	yes	*	*		*					
amitriptyline	nortriptyline	yes						*			
amitriptyline	amlodipine	no			*						
amlodipine	amiloride	yes	*	*		*					
amlodipine	amiodarone	no				*					
anagrelide	anastrozole	no				*					
atenolol	tenofovir	no			*						
atomoxetine	atorvastatin	yes	*								
atorvastatin	atomoxetine	yes	*								

azacitidine	azathioprine	yes	*				
azathioprine	azacitidine	yes	*				
azathioprine	azithromycin	yes				*	*
azithromycin	azathioprine	yes				*	*
buprenorphine	hydromorphone	yes	*				
bupropion	buspirone	yes	*	*		*	
buspirone	bupropion	yes	*	*		*	
buspirone	risperidone	no				*	
captopril	carvedilol	yes	*	*			
carbamazepine	oxcarbazepine	yes	*				
carbidopa	carvedilol	no				*	
carvedilol	captopril	yes	*	*			
cefuroxime	sulfasalazine	yes	*				
cetirizine	sertraline	yes	*				
chlordiazepoxide	chlorpromazine	yes	*			*	
chlorpromazine	chlordiazepoxide	yes	*			*	
chlorpromazine	clomipramine	no				*	
cisplatin	carboplatin	no	*	*			
clonazepam	clonidine	yes	*	*		*	
clonazepam	clozapine	yes	*			*	
clonazepam	lorazepam	yes	*	*		*	

| | | |

| | | |

hydrochlorothiazide	hydroxyzine	yes	*		*		*	
hydrocortisone	hydralazine	no				*		
hydrocortisone	hydrochlorothiazide	no			*	*		
hydromorphone	buprenorphine	yes	*					
hydromorphone	haloperidol	no				*		
hydromorphone	hydralazine	no			*			
hydroxychloroquine	hydrochlorothiazide	no				*		
hydroxyzine	hydralazine	yes	*	*	*		*	
hydroxyzine	hydrochlorothiazide	yes	*		*		*	
indapamide	isosorbide	no				*		
infliximab	rituximab	no	*					
isosorbide	isotretinoin	no				*		
isotretinoin	tretinoin	yes	*			*		
ketorolac	metoprolol	no			*			
labetalol	lamotrigine	yes	*					
lamivudine	lamotrigine	yes	*			*		
lamotrigine	labetalol	yes	*					
lamotrigine	lamivudine	yes	*			*		
lamotrigine	levetiracetam	yes	*			*		
levetiracetam	lamotrigine	yes	*			*		
levetiracetam	levocarnitine	yes	*			*		

levetiracetam	levofloxacin	yes	*							
levocarnitine	levetiracetam	yes	*				*			
levocarnitine	levofloxacin	yes						*		
levofloxacin	levetiracetam	yes	*							
levofloxacin	levocarnitine	yes						*		
levofloxacin	levodopa	no					*			
linaclotide	linagliptin	no	*							
lisinopril	verapamil	no					*			
loperamide	lorazepam	no					*			
lorazepam	alprazolam	yes	*	*						
lorazepam	clonazepam	yes	*	*			*			
lorazepam	loratadine	no					*			
lorazepam	oxazepam	no					*			
losartan	lisinopril	no					*			
medroxyprogesterone	methylprednisolone	yes	*	*						
metformin	metronidazole	yes	*		*					
metformin	methocarbamol	no					*			
methenamine	methionine	no							*	
methocarbamol	metronidazole	no					*			
methotrexate	minoxidil	no					*			
methylprednisolone	medroxyprogesterone	yes	*	*						

methylprednisolone	methyldopa	no			*				
metoclopramide	metoprolol	yes			*		*		
metoclopramide	metronidazole	no				*			
metoprolol	metoclopramide	yes			*		*		
metoprolol	metformin	no			*				
metoprolol	mirtazapine	no				*			
metronidazole	metformin	yes	*		*				
metronidazole	metoprolol	no			*				
mifepristone	misoprostol	yes	*			*			
minoxidil	midodrine	no				*			
mirtazapine	temazepam	no				*			
misoprostol	mifepristone	yes	*			*			
mitomycin	mitoxantrone	yes	*	*					
mitoxantrone	mitomycin	yes	*	*					
nadolol	haloperidol	no				*			
nicardipine	nifedipine	yes	*						
nicardipine	nimodipine	no	*						
nicotinamide	nicotinic acid	yes						*	
nicotinic acid	nicotinamide	yes						*	
nifedipine	nicardipine	yes	*						
nifedipine	nimodipine	no	*						

nilotinib	niraparib	no	*						
nortriptyline	amitriptyline	yes					*		
nortriptyline	nifedipine	no				*			
olanzapine	quetiapine	yes	*						
olanzapine	sertraline	no				*			
olsalazine	olanzapine	no				*			
oxcarbazepine	carbamazepine	yes	*						
paroxetine	duloxetine	yes	*						
paroxetine	fluoxetine	yes	*						
paroxetine	piroxicam	yes	*						
piroxicam	paroxetine	yes	*						
progesterone	dinoprostone	no							*
promazine	promethazine	yes					*		
promethazine	promazine	yes					*		
pyridoxine	paroxetine	no			*				
pyridoxine	primidone	no				*			
quetiapine	olanzapine	yes	*						
rabeprazole	donepezil	no				*			
raloxifene	rabeprazole	no				*			
ranitidine	clonidine	no			*				
ribavirin	riboflavin	no	*						

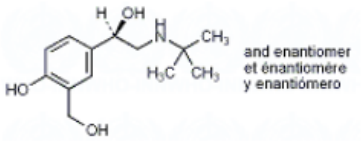
risperidone	ropinirole	yes	*					
ropinirole	risperidone	yes	*					
ropinirole	repaglinide	no			*			
saxagliptin	sitagliptin	yes	*					
sertraline	cetirizine	yes	*					
simvastatin	valsartan	no			*			
sitagliptin	saxagliptin	yes	*					
sitagliptin	sumatriptan	yes	*		*			
spironolactone	sertraline	no			*			
sufentanil	fentanyl	yes	*					*
sulfadiazine	sulfasalazine	yes	*			*		
sulfasalazine	cefuroxime	yes	*					
sulfasalazine	sulfadiazine	yes	*			*		
sumatriptan	sitagliptin	yes	*		*			
sumatriptan	zolmitriptan	yes	*	*				
sumatriptan	somatropin	no				*		
tacrolimus	tamsulosin	yes	*		*			
tamsulosin	tacrolimus	yes	*		*			
temazepam	diazepam	no			*			
tenofovir	trandolapril	no			*			
terbutaline	terazosin	no			*			

tetracycline	sertraline	no			*	
tolterodine	tacrolimus	no				*
tramadol	trazodone	yes	*			*
trazodone	tramadol	yes	*			*
tretinoin	isotretinoin	yes	*			*
triamterene	trimethoprim	no				*
trihexyphenidyl	ramipril	no				*
vigabatrin	dabigatran	yes	*			
vinblastine	vincristine	yes	*	*		
vincristine	vinblastine	yes	*	*		
zolmitriptan	sumatriptan	yes	*	*		

A7 MedNet listing of salbutamol and its translations

INN

salbutamol



Latin	salbutamolum
French	salbutamol
Spanish	salbutamol
Russian	сальбутамол
Arabic	سالبوتامول
Chinese	沙丁胺醇
Phonetic	salbutamol
Molecular formula	$C_{13}H_{21}NO_3$
ATC Codes	R03AC02
	R03CC02

National commission(s)	**Alternate name(s)**
DCIT	salbutamolo
JAN	salbutamol sulfate
Ph. Int.	salbutamoli sulfas
RS	salbutamol sulfate
USAN	albuterol
USAN	albuterol sulfate
USP	albuterol

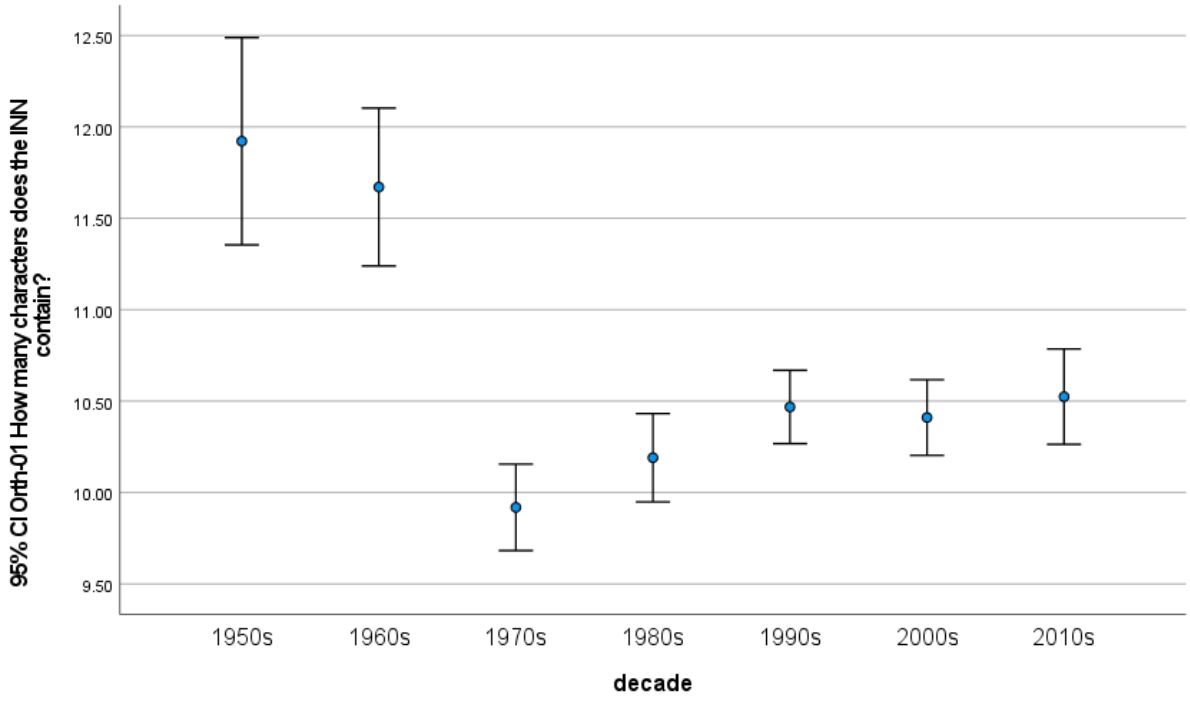
National commission(s) using same name

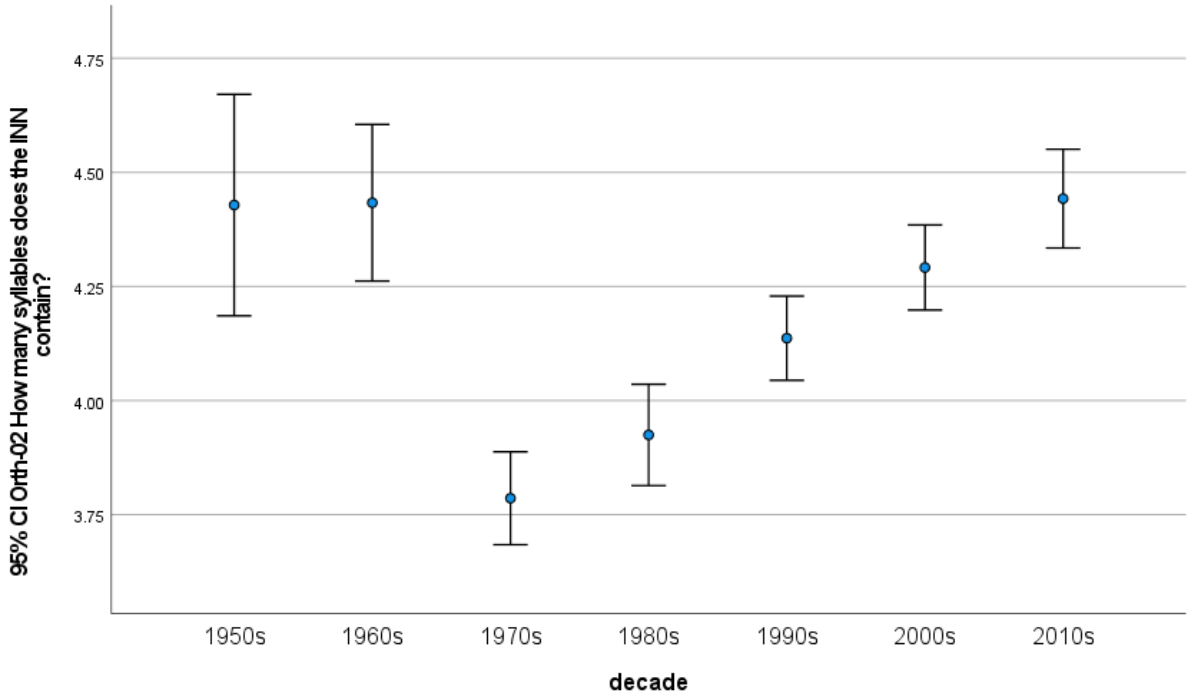
BAN
BPC
ChP
CSP
DCF
FEUM
IP
JP
Ph. Eur.
Ph. Int.
RS

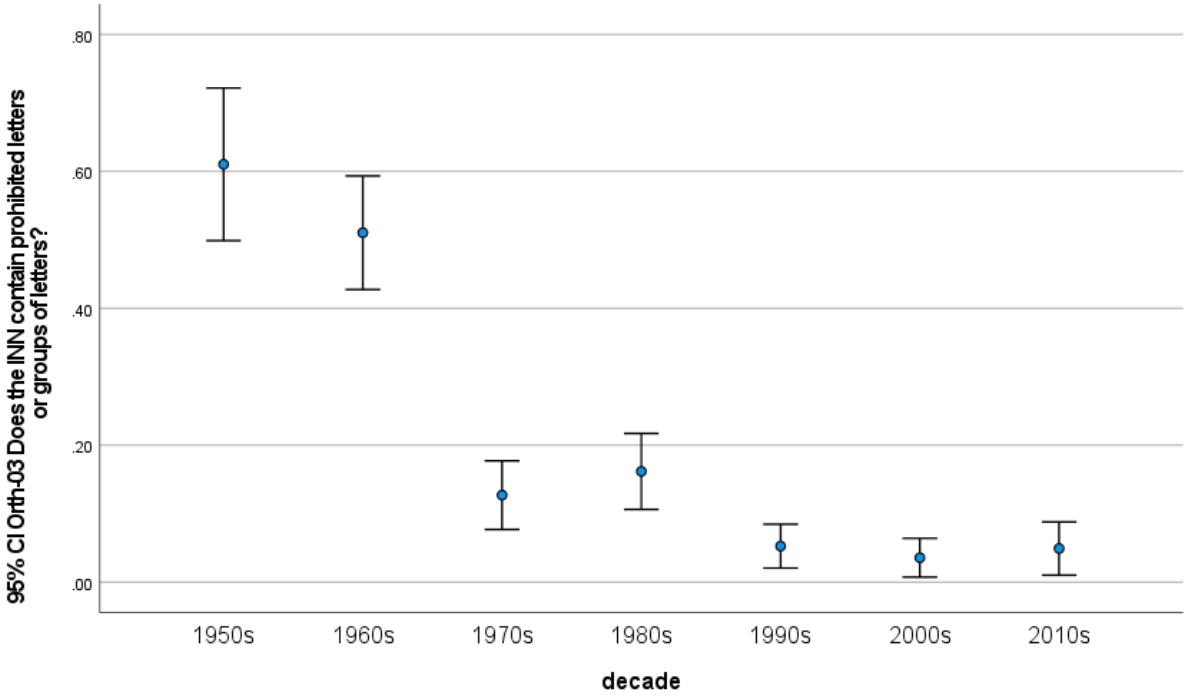
A8 Methods to analyse formal measures

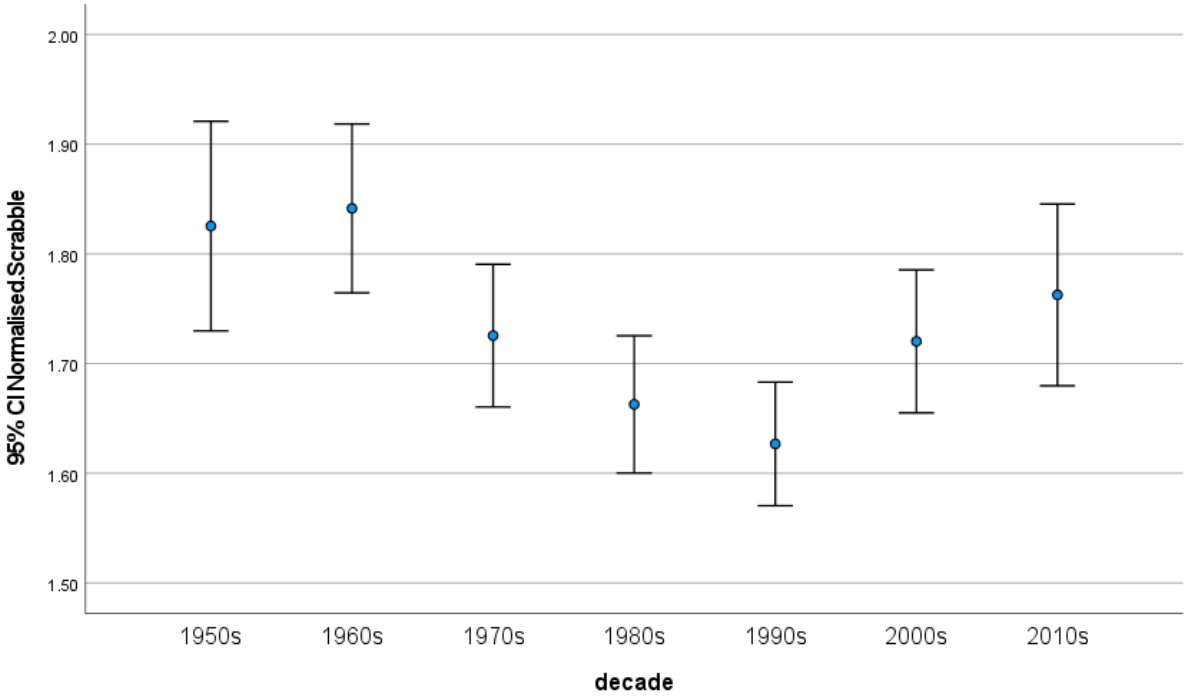
Measures	Methods
Number of words, characters, and syllables (1.1; 1.2; and 1.3)	Number of words and number of letters were calculated using Excel functions. In English, the number of syllables in a word roughly corresponds to the number of vowels, and in this case it was calculated using a method similar to that described by Lambert (Lambert et al., 2001a): the number of vowels (including y), less a final -e. Furthermore, combinations of double or triple vowels that form a single vowel sound or form a diphthong were discounted - a diphthong is a one-syllable sound formed by combining two vowel sounds, such as coin, loud and side, so these vowels should be counted as a single vowel. This was calculated using COUNTIF, LEFT, and MID functions. This formula was able to predict the number of syllables with 90% accuracy. Since only 160 names contained double or triple vowels, each was looked at 'by hand' to ascertain whether the double vowel was a single vowel syllable, a diphthong, or two separate syllables. Then the number of syllables was adjusted accordingly. For example, acetylcysteine has seven vowels (including y): less the final -e, and less the double vowel (-ei-) that is pronounced as a single sound (phoneme), resulting in five syllables.
Common beginnings and endings (1.4)	Beginnings and endings of names were spliced into two-letter strings (bigrams, i.e. two letter ngrams) and three-letter strings (trigrams, i.e. three-letter ngrams), using LEFT and RIGHT trim functions. These were ranked by frequency and proportion of total ngrams in that position.
Orthographic depth (1.5)	The WHO INN programme has prohibited many letters or groups of letters with the intention of making INNs easy to use internationally. These are letters that are more likely to be pronounced differently in languages other than English (e.g. <th> and <ph>), or letters that may not be present in other languages (e.g. <k>, <h>, or <y>). This is guideline 7 in the INN guidelines, and the rule was recently updated to further prohibit the use of duplicate letters. The guidance states: "To facilitate the translation and pronunciation of INN, 'f' should be used instead of 'ph', 't' instead of 'th', 'e' instead of 'ae' or ''oe'', and 'i' instead of 'y'; the use of the letters 'h' and 'k' should be avoided. When devising an INN it is important to be aware of possible language problems. Since the name is used worldwide, not only should certain letters be avoided, but experts need to be aware of unsuitable connotations in the major languages spoken in the world." To measure orthographic depth, wildcard searches were used to identify the presence and frequency of prohibited letters and duplicate letters (e.g. 'pp') in INNS.

A9 Orthographic measures, over time





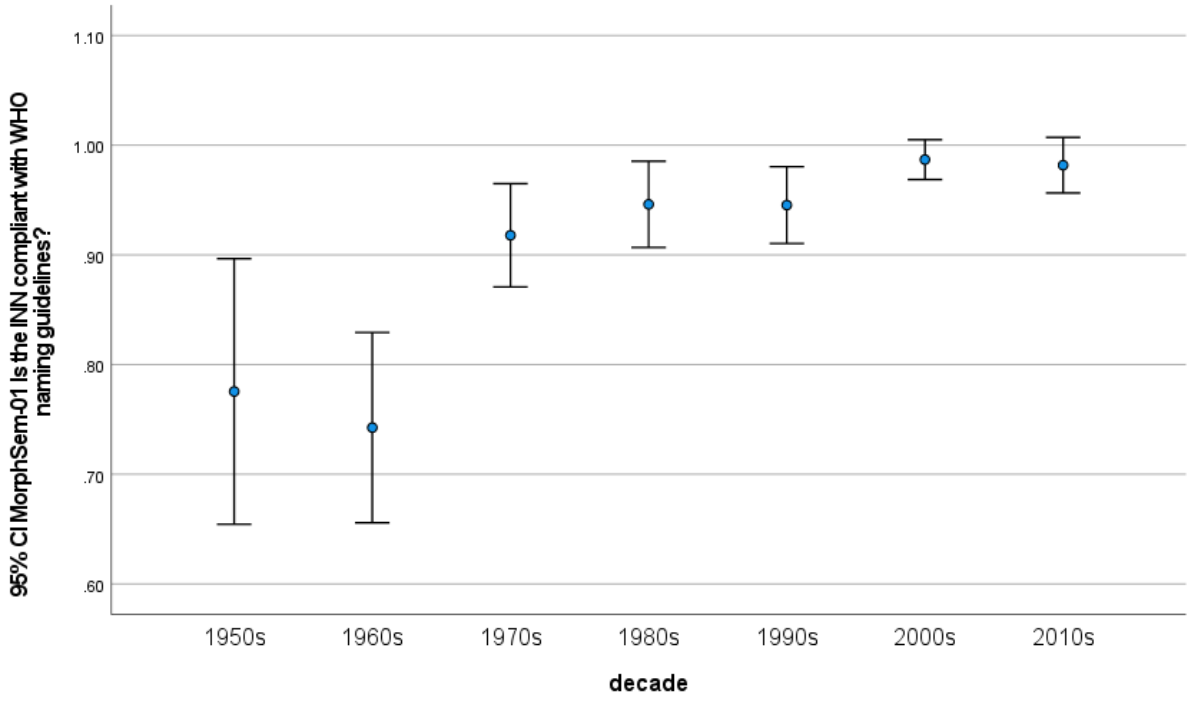

95% CI Orth-03 Does the INN contain prohibited letters or groups of letters?
.80
.60
.40
.20
.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

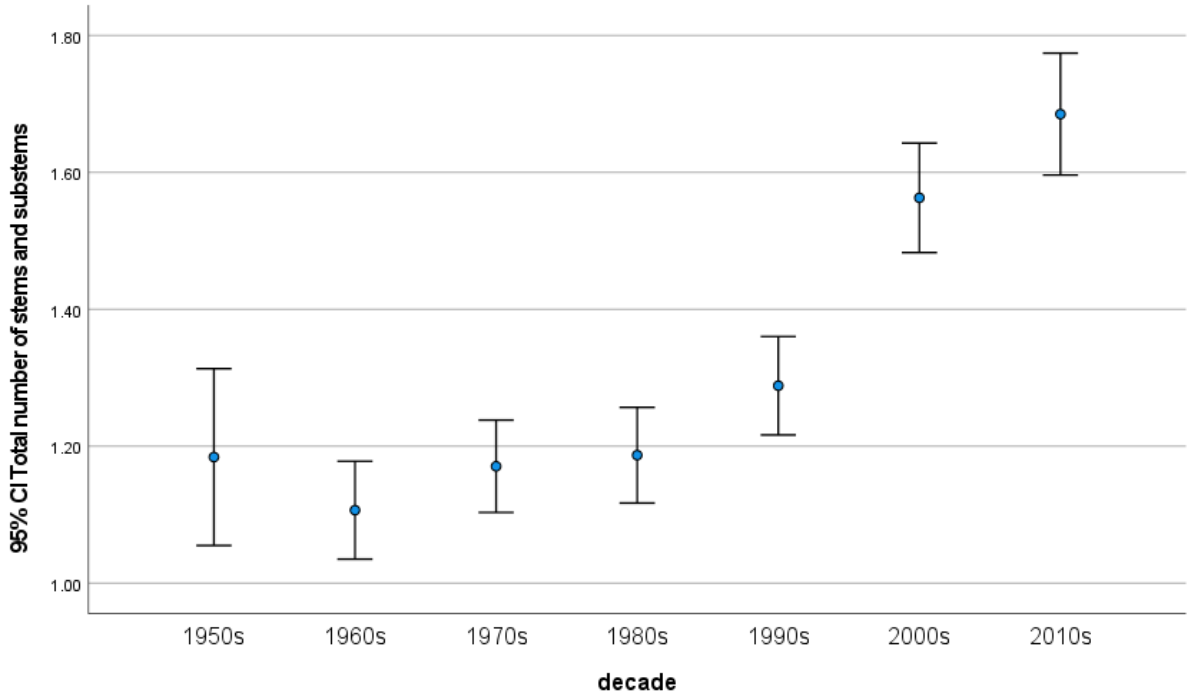

95% CI Normalised.Scrabble
2.00
1.90
1.80
1.70
1.60
1.50
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

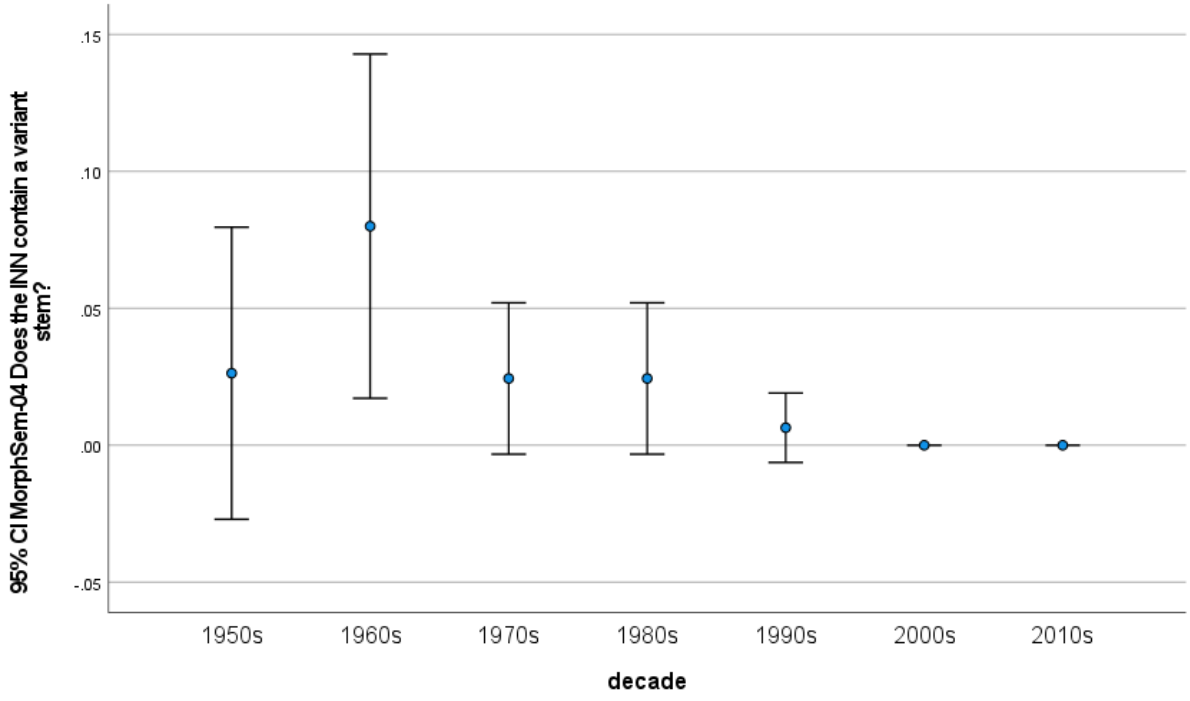
A10 INNs using a variant STEM

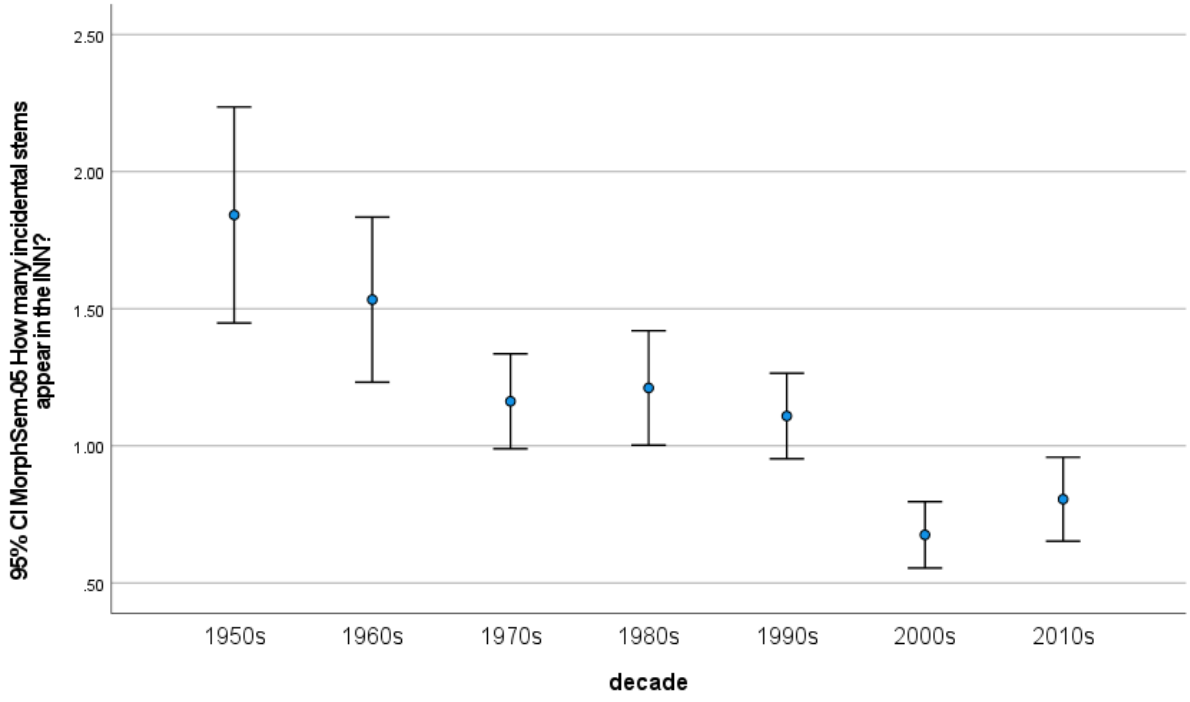
INN	Official STEM	Variant STEM
alfuzosin	azosin	uzosin
cyclophosphamide	fos	phosphamide
doxepin	pine	pin
epinephrine	drine	phrine
fentanyl	fentanil	fentanyl
hydroxyzine	izine	yzine
nicardipine	nico	nic
nitrazepam	azepam	nitr
norepinephrine	drine	phrine
phenylephrine	drine	phrine
ranitidine	tidine	nit
tamoxifen	ifene	oxifen
tamsulosin	azosin	ulosin

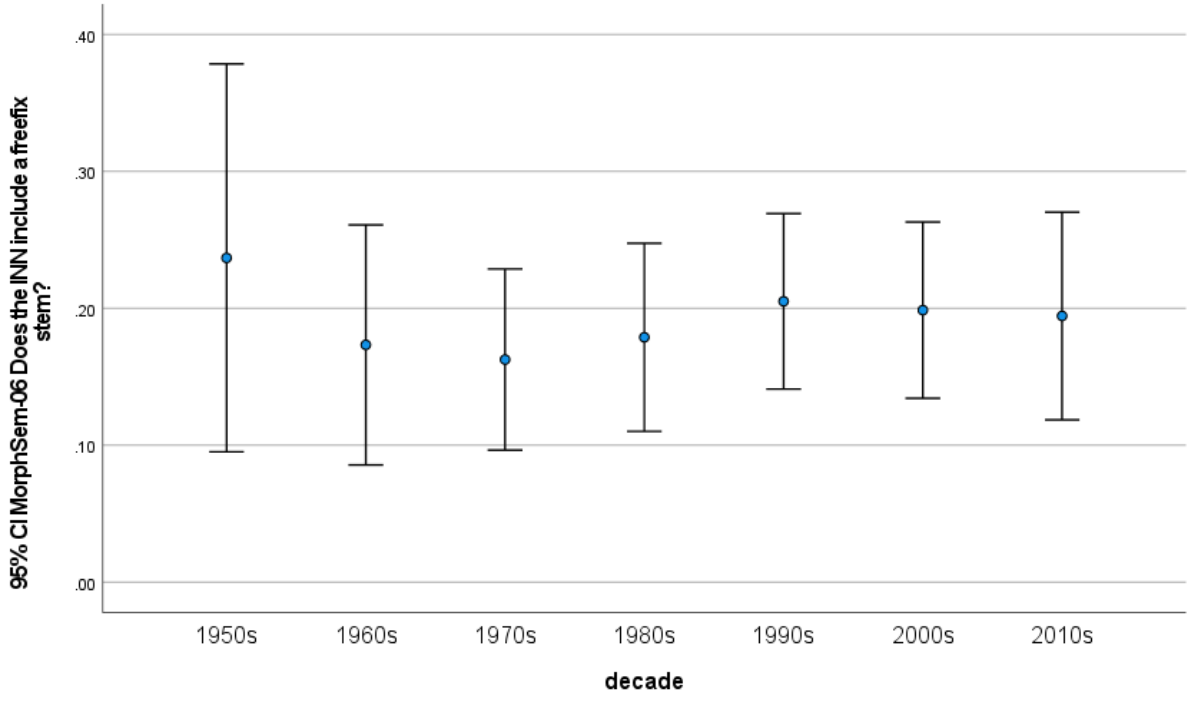
A11 Morphosemantic measures, over time

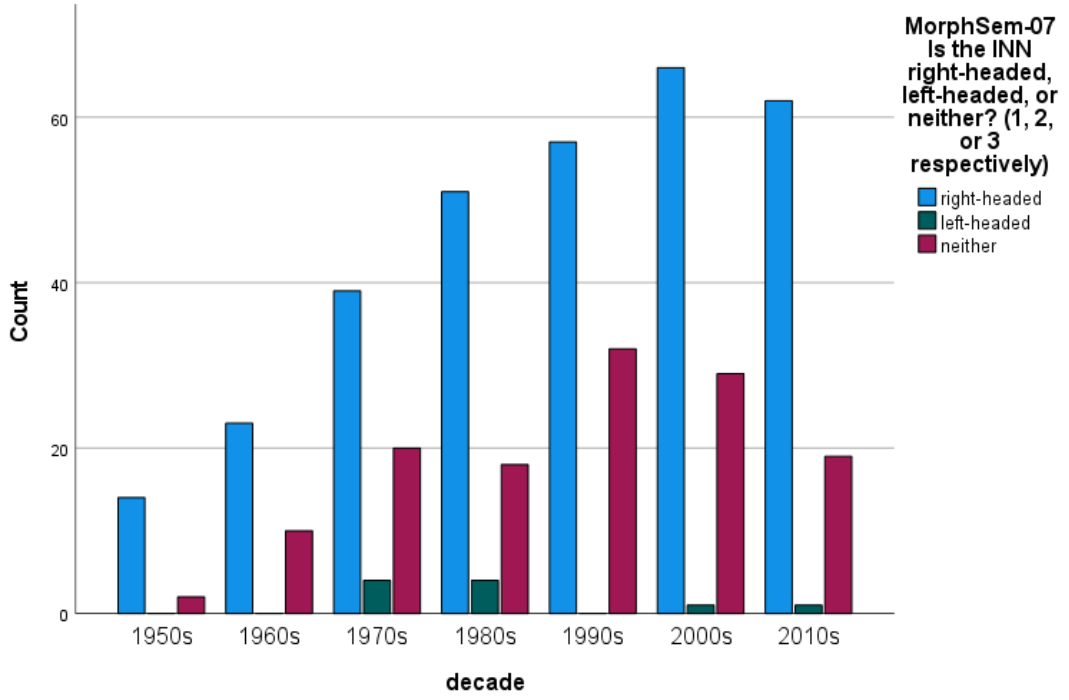


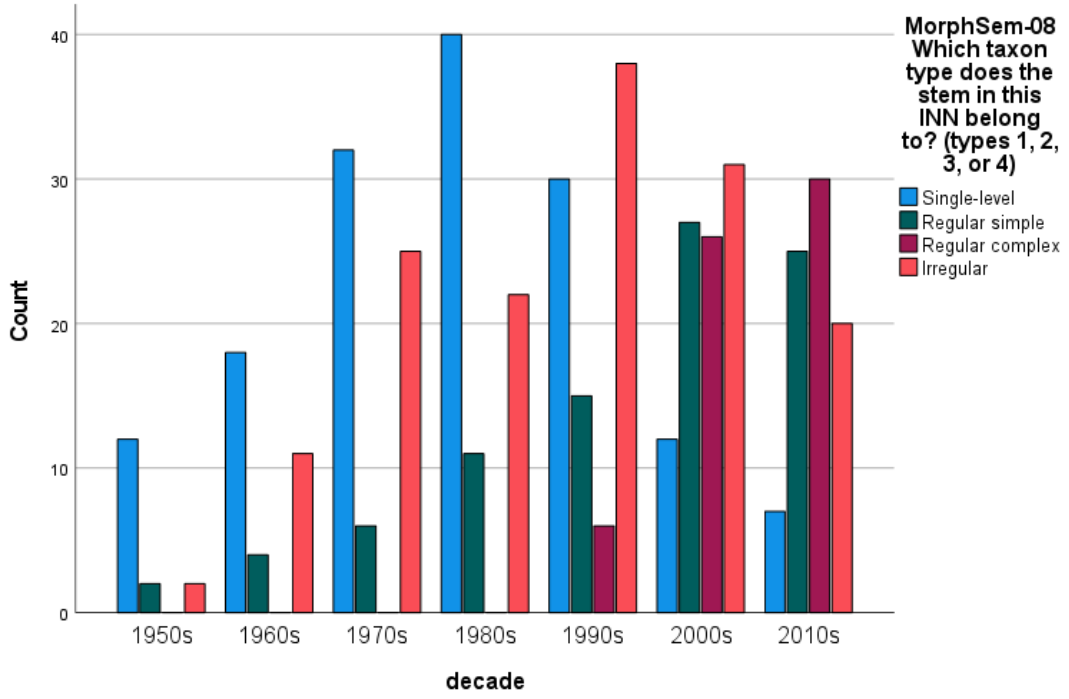



95% CI MorphSem-04 Does the INN contain a variant stem?
.15
.10
.05
.00
-.05
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

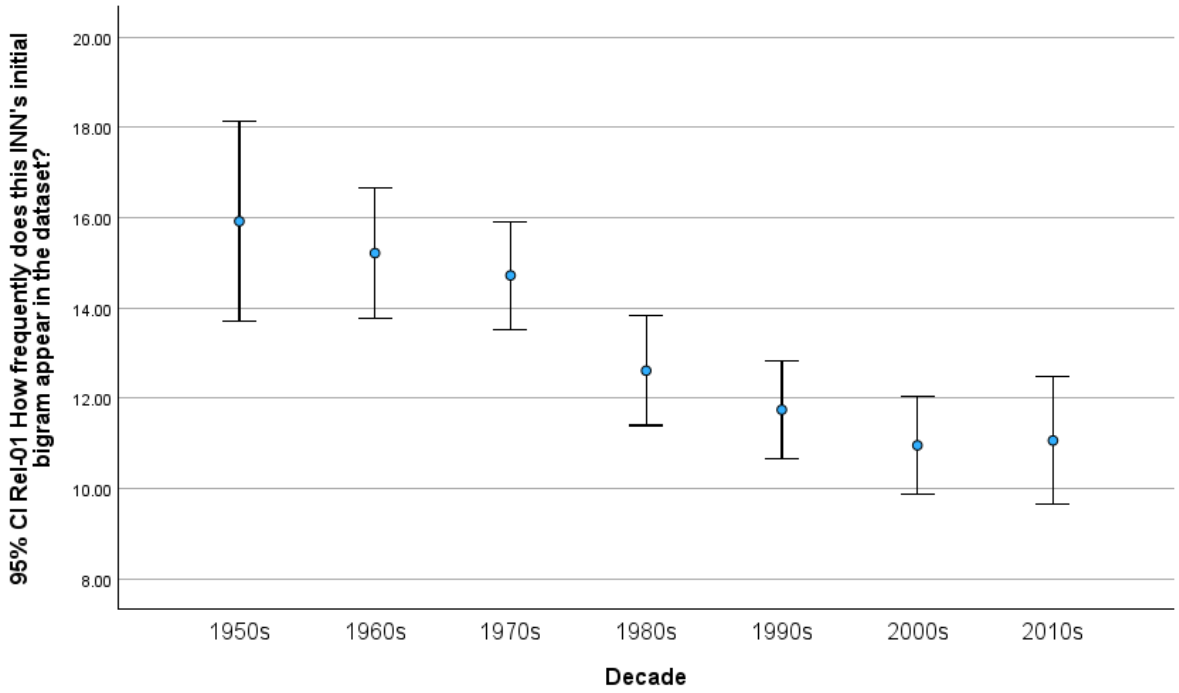

95% CI MorphSem-05 How many incidental stems appear in the INN?
2.50
2.00
1.50
1.00
.50
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

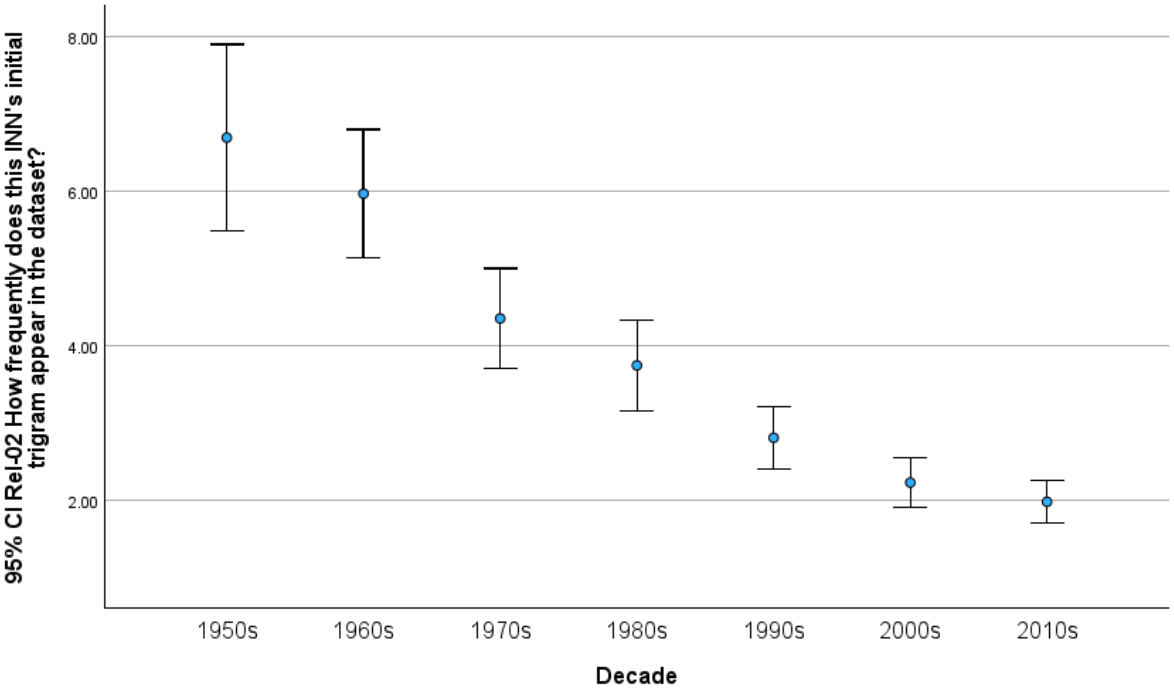

95% CI MorphSem-06 Does the INN include a freefix stem?
.40
.30
.20
.10
.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

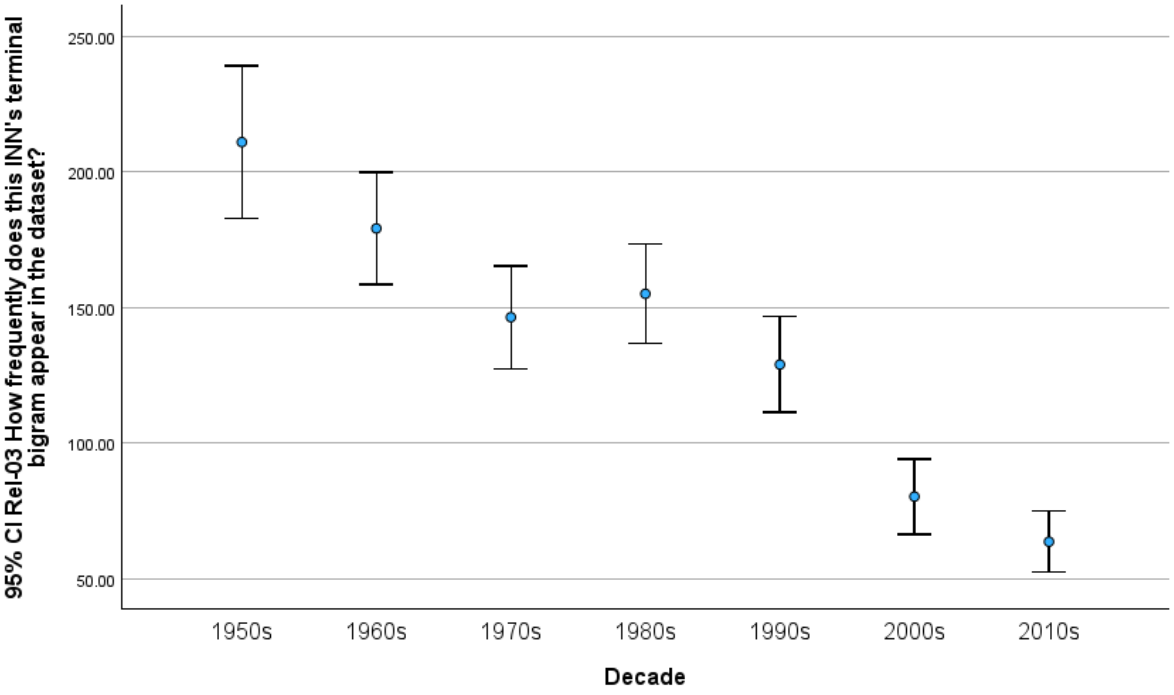

Count
60
40
20
0
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade
MorphSem-07 Is the INN right-headed, left-headed, or neither? (1, 2, or 3 respectively)
right-headed
left-headed
neither

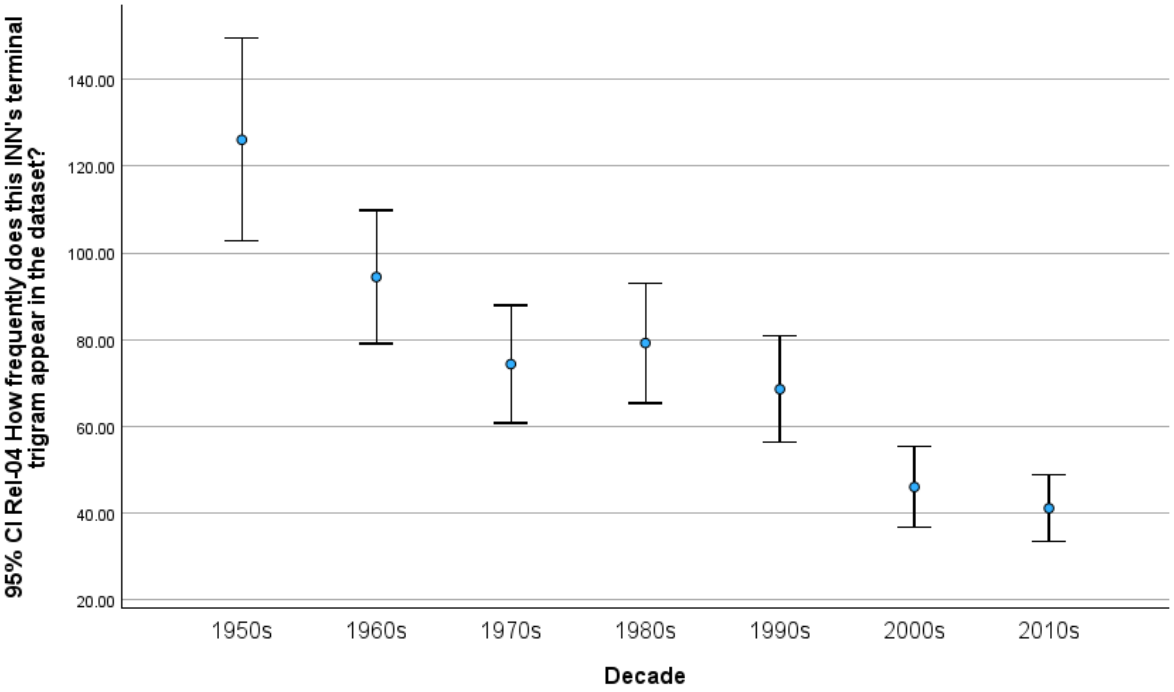


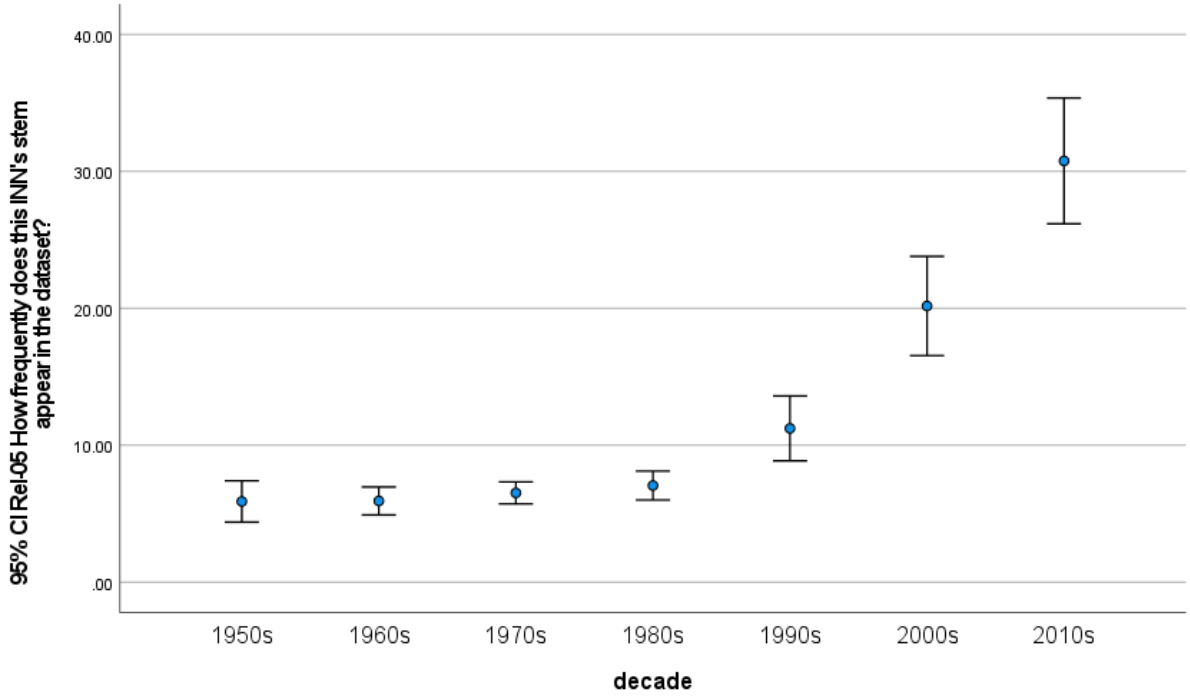
A12 Relational measures, over time

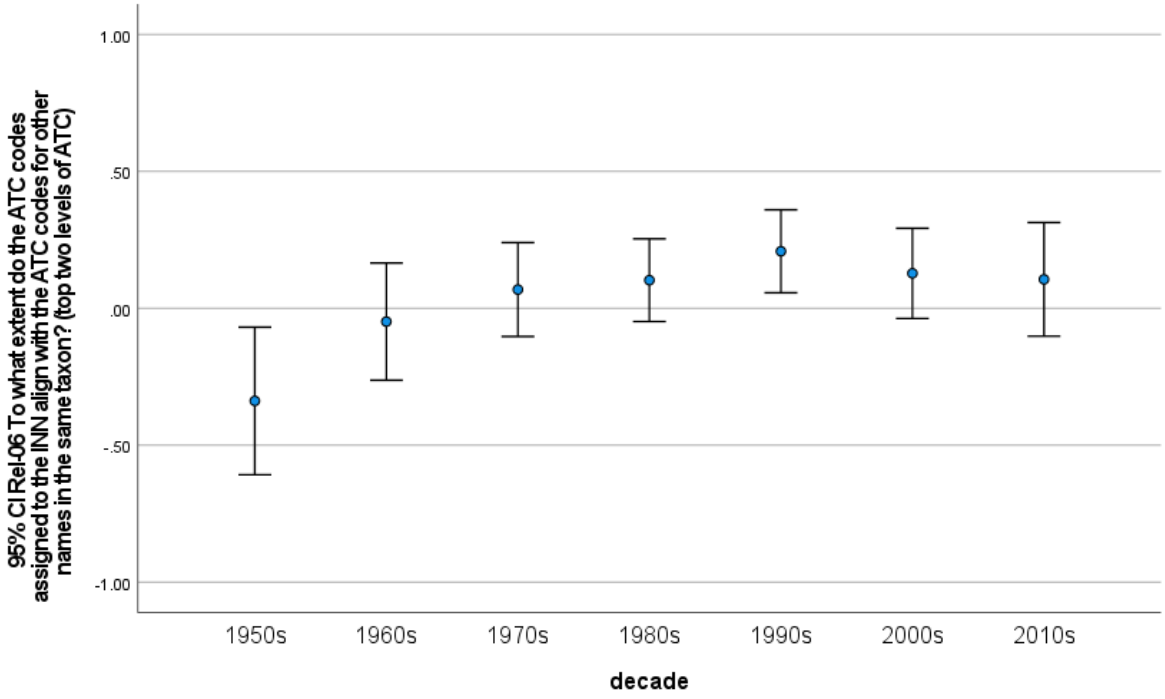


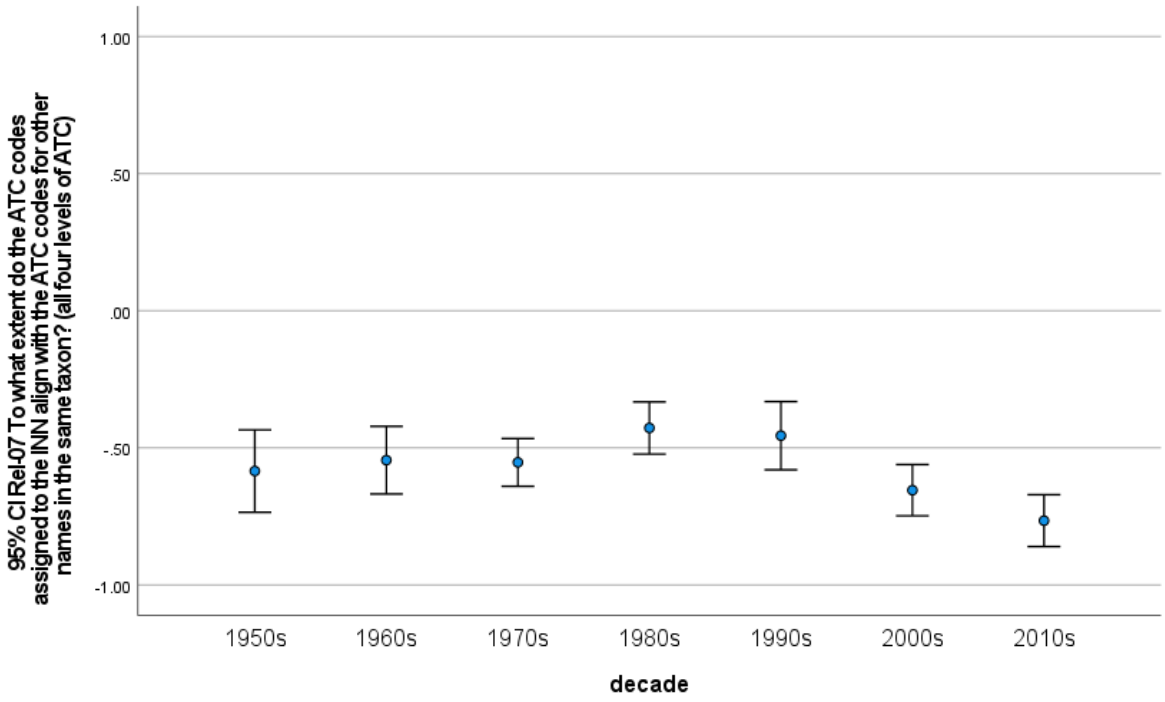

95% CI Rel-02 How frequently does this INN's initial trigram appear in the dataset?
8.00
6.00
4.00
2.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Decade

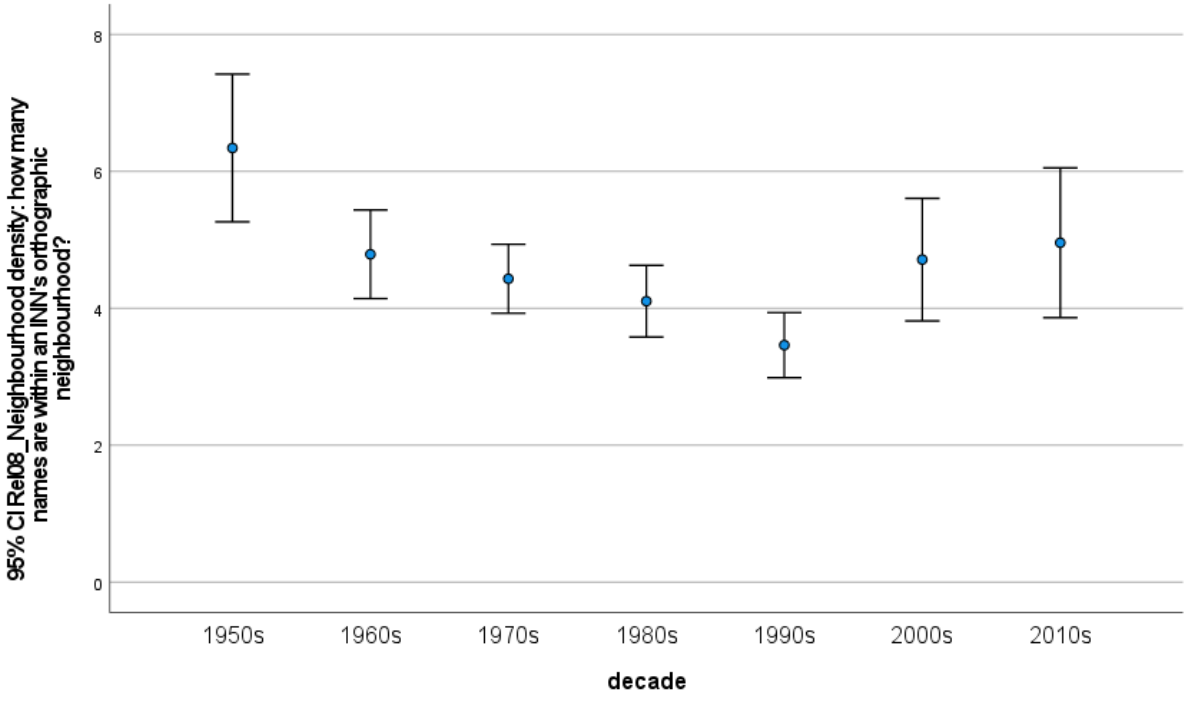

95% CI Rel-03 How frequently does this INN's terminal bigram appear in the dataset?
250.00
200.00
150.00
100.00
50.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Decade

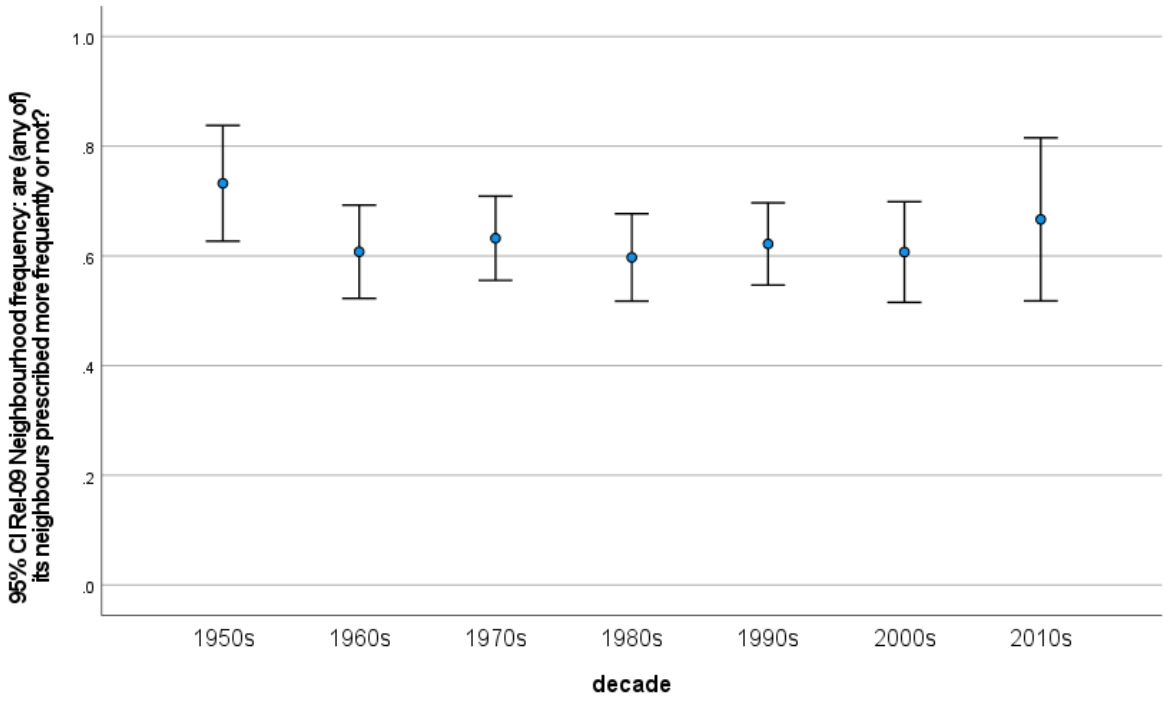

95% CI Rel-04 How frequently does this INN's terminal trigram appear in the dataset?
140.00
120.00
100.00
80.00
60.00
40.00
20.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Decade

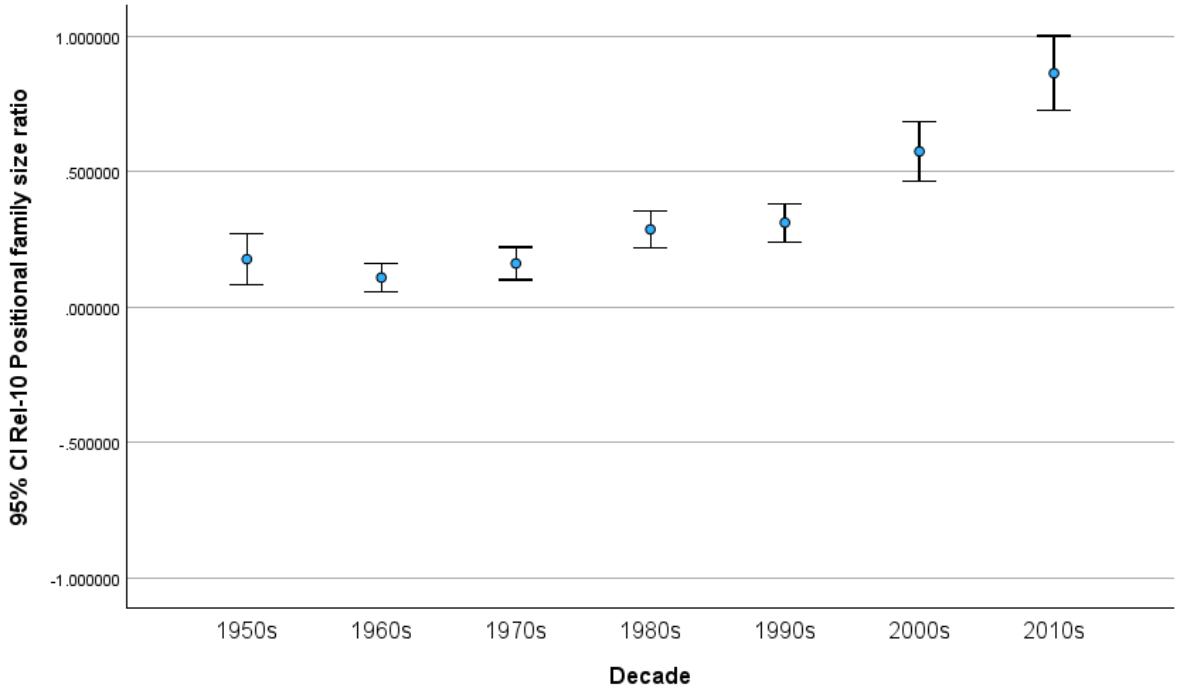

95% CI Rel-05 How frequently does this INN's stem appear in the dataset?
40.00
30.00
20.00
10.00
.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

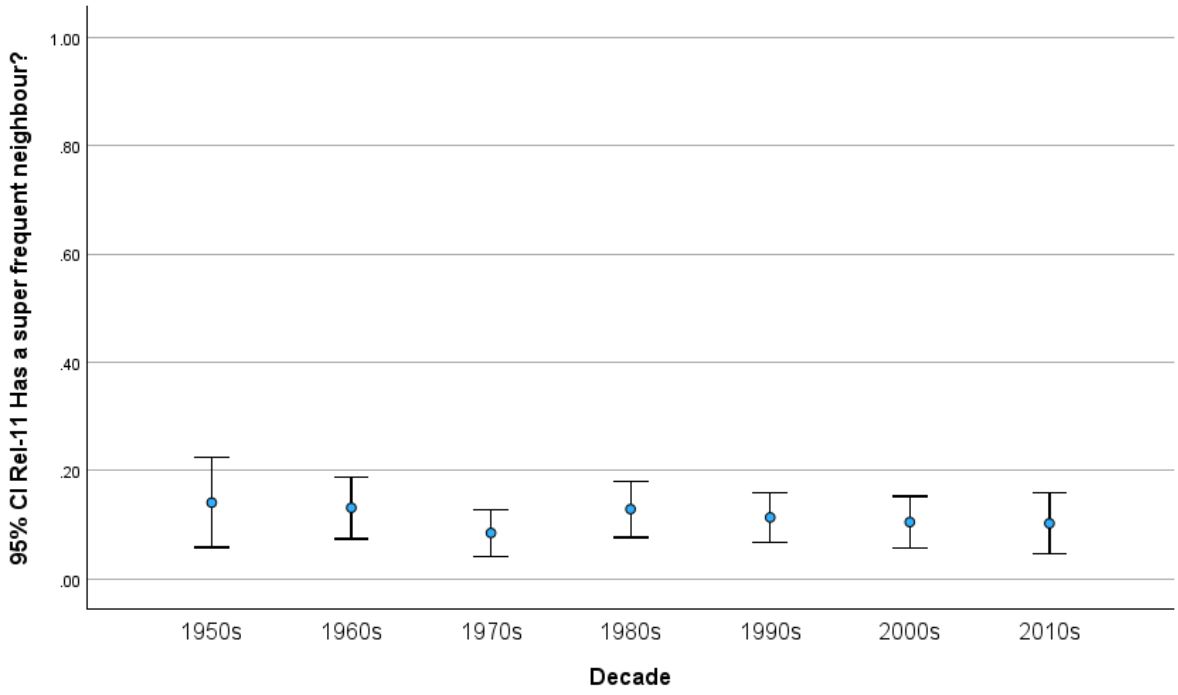

95% CI Rel-06 To what extent do the ATC codes assigned to the INN align with the ATC codes for other names in the same taxon? (top two levels of ATC)
1.00
.50
.00
-.50
-1.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade


95% CI Rel-07 To what extent do the ATC codes assigned to the INN align with the ATC codes for other names in the same taxon? (all four levels of ATC)
1.00
.50
.00
-.50
-1.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade


95% CI Rel08_Neighbourhood density: how many names are within an INN's orthographic neighbourhood?
8
6
4
2
0
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade

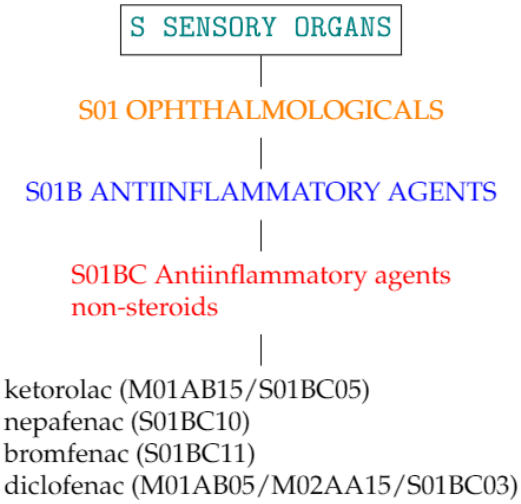

95% CI Rel-09 Neighbourhood frequency: are (any of) its neighbours prescribed more frequently or not?
1.0
.8
.6
.4
.2
.0
1950s
1960s
1970s
1980s
1990s
2000s
2010s
decade


95% CI Rel-10 Positional family size ratio
1.000000
.500000
.000000
-.500000
-1.000000
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Decade

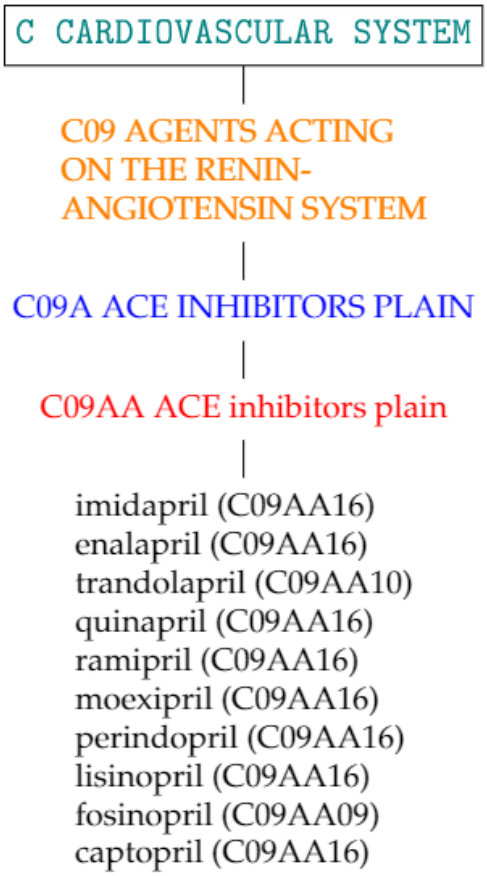

95% CI Rel-11 Has a super frequent neighbour?
1.00
.80
.60
.40
.20
.00
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Decade

A13 Examples of ATC codes for names in four taxa

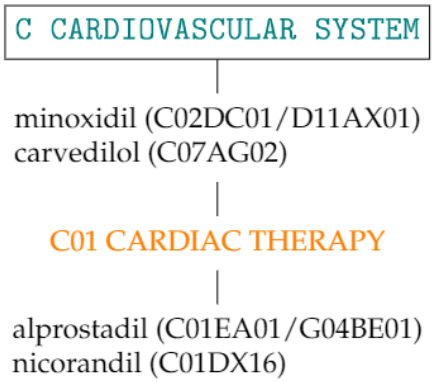
a: -ac taxon



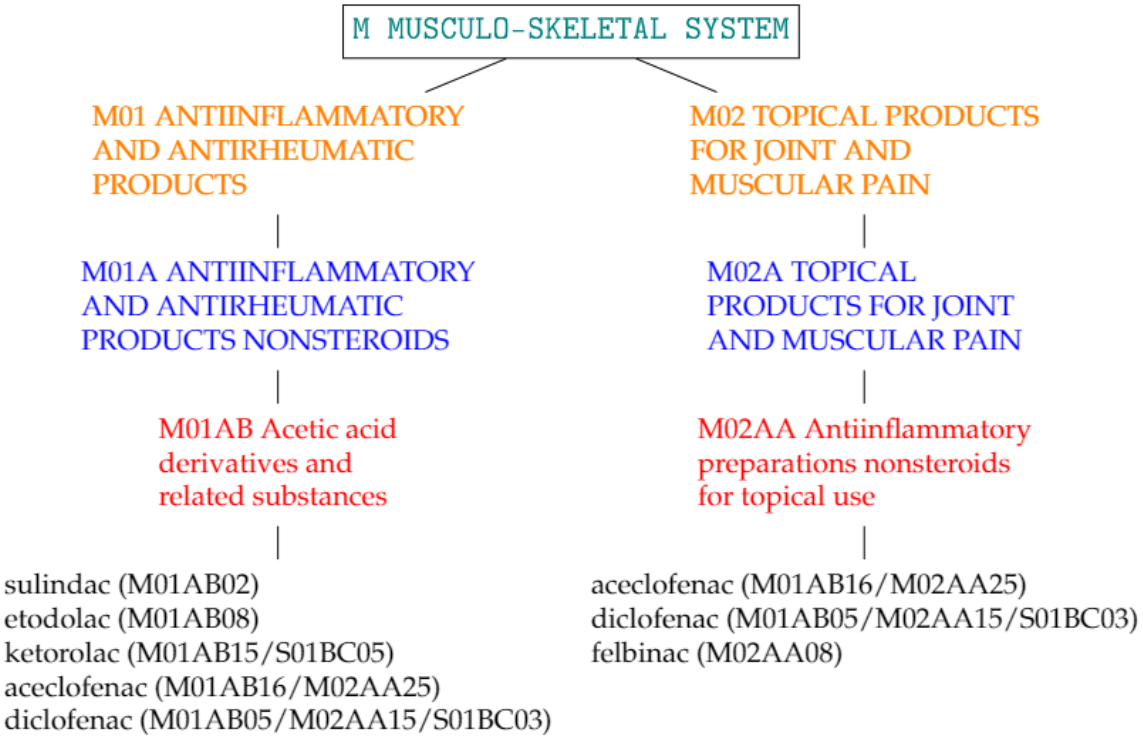
b: -pril taxon



c: -dil taxon



d: -ac taxon



A14 STEMS used in INNs with a high number of neighbours

Stems used in INNs with a high number of neighbours

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	mab	21	45.7	55.3	55.3
	tinib	11	23.9	28.9	84.2
	caine	2	4.3	5.3	89.5
	clidine	1	2.2	2.6	92.1
	olol	1	2.2	2.6	94.7
	pamide	1	2.2	2.6	97.4
	prost	1	2.2	2.6	100.0
	Total	38	82.6	100.0	
Missing	-999.00	8	17.4		
Total		46	100.0		

A15 All STEM taxa in English, Russian, and Italian nomenclatures[51]

English-language INNs			Russian-language INNs			Italian-language DCITs			Pharmacological group (WHO definition)
STEM	SUBSTEM	Names	STEM	SUBSTEM	Names	STEM	SUBSTEM	Names	--
abine	arabine	clofarabine cytarabine fludarabine nelarabine	абин	арабин	клофарабин цитарабин флударабин неларабин	abina	arabina	clofarabina citarabina fludarabina nelarabina	arabinofuranosyl derivatives / nucleosides antiviral or antineoplastic agents, cytarabine or azacitidine derivatives
	citabine	capecitabine decitabine emtricitabine gemcitabine		цитабин	капецитабин децитабин эмтрицитабин гемцитабин		citabina	capecitabina decitabina emtricitabina gemcitabina	
ac	[no SUBSTEM]	felbinac ketorolac sulindac	ак	[no SUBSTEM]	фелбинак кеторолак сулиндак	ac	[no SUBSTEM]	felbinac ketorolac sulindac	anti-inflammatory agents, ibufenac derivatives
	clofenac	aceclofenac diclofenac		клофенак	ацеклофенак диклофенак		clofenac	aceclofenac diclofenac	
	dolac	etodolac		долак	этодолак		dolac	etodolac	
	fenac	bromfenac nepafenac		фенак	бромфенак непафенак		fenac	bromfenac nepafenac	
actide	[no SUBSTEM]	tetracosactide	актид	[no SUBSTEM]	тетракозактид	actide	[no SUBSTEM]	tetracosactide	synthetic polypeptide with a corticotropin-like action
adol	[no SUBSTEM]	tapentadol tramadol	адол	[no SUBSTEM]	тапентадол трамадол	adol	[no SUBSTEM]	tapentadol tramadolo	analgesics

[51] The font size is very small for Appendix section A14 as only one appendix file could be submitted, so it is in a standard A4 size with the rest of the tables. However, with the original submission a PDF file was also submitted, on which can be zoomed in to consult the table.

	adoline	eluxadoline		адолин	элуксадолин		adolina	eluxadolina	
afenone	[no SUBSTEM]	propafenone	афенон	[no SUBSTEM]	пропафенон	afenone	[no SUBSTEM]	propafenone	antiarrhythmics, propafenone derivatives
afil	[no SUBSTEM]	avanafil sildenafil tadalafil vardenafil	афил	[no SUBSTEM]	аванафил силденафил тадалафил варденафил	afil	[no SUBSTEM]	avanafil sildenafil tadalafil vardenafil	inhibitors of phosphodiesterase PDE5 with vasodilator action
ampanel	[no SUBSTEM]	perampanel	ампанел	[no SUBSTEM]	перампанел	ampanel	[no SUBSTEM]	perampanel	antagonists of the ionotropic non-NMDA (N-methyl-d-aspartate) glutamate receptors (Namely the AMPA (amino-hydroxymethylisoxazole-propionic acid) and/or KA (kainite antagonist) receptors)
andr	[no SUBSTEM]	nandrolone	андр	[no SUBSTEM]	нандролон	andr	[no SUBSTEM]	nandrolone	steroids, androgens
	sterone	prasterone testosterone		стерон	прастерон тестостерон		sterone	prasterone testosterone	
anib	[no SUBSTEM]	nintedanib pazopanib tivozanib vandetanib	аниб	[no SUBSTEM]	нинтеданиб пазопаниб тивозаниб вандетаниб	anib	[no SUBSTEM]	nintedanib pazopanib tivozanib vandetanib	angiogenesis inhibitors
anide	etanide	bumetanide	анид	етанид	буметанид	anide	etanide	bumetanide	saluretics
antrone	[no SUBSTEM]	mitoxantrone pixantrone	антрон	[no SUBSTEM]	митоксантрон пиксантрон	antrone	[no SUBSTEM]	mitoxantrone pixantrone	antineoplastics; anthraquinone derivatives
arol	[no SUBSTEM]	acenocoumarol	арол	[no SUBSTEM]	аценокумарол	arol	[no SUBSTEM]	acenocumarolo	anticoagulants, dicoumarol derivatives
arone	[no SUBSTEM]	amiodarone	арон	[no SUBSTEM]	амиодарон дронедарон	arone	[no SUBSTEM]	amiodarone	antiarrhythmics

		dronedarone						dronedarone	
arotene	[no SUBSTEM]	bexarotene tazarotene	аротен	[no SUBSTEM]	бексаротен тазаротен	arotene	[no SUBSTEM]	bexarotene tazarotene	arotinoid derivatives
ase	[no SUBSTEM]	cocarboxylase hyaluronidase laronidase pegaspargase rasburicase streptokinase urokinase	аза	[no SUBSTEM]	кокарбоксилаза гиалуронидаза ларонидаза пэгаспаргаза разбуриказа стрептокиназа урокиназа	asi	[no SUBSTEM]	cocarbossilasi ialuronidasi laronidasi pegaspargasi rasburicasi streptochinasi urochinasi	enzymes
	glucerase	imiglucerase		глюцераза	имиглюцераза		glucerasi	imiglucerasi	
	sulfase	idursulfase		сульфаза	идурсульфаза		sulfasi	idursulfasi	
	teplase	alteplase reteplase tenecteplase		теплаза	алтеплаза ретеплаза тенектеплаза		teplasi	alteplasi reteplasi tenecteplasi	
ast	lukast	montelukast	аст	лукаст	монтелукаст	ast	lukast	montelukast	anti-allergic or anti-inflammatory, not acting as anti-histaminics
	milast	apremilast		миласт	апремиласт		milast	apremilast	
		roflumilast			рофлумиласт			roflumilast	
astine	[no SUBSTEM]	acrivastine azelastine bilastine clemastine epinastine mizolastine	астин	[no SUBSTEM]	акривастин азеластин биластин клемастин эпинастин мизоластин	astina	[no SUBSTEM]	acrivastina azelastina bilastina clemastina epinastine mizolastina	antihistaminics

azenil	[no SUBSTEM]	flumazenil	азенил	[no SUBSTEM]	флумазенил	azenil	[no SUBSTEM]	flumazenil	benzodiazepine receptor antagonists/agonists (benzodiazepine derivatives)
azepam	[no SUBSTEM]	clonazepam diazepam flurazepam lorazepam lormetazepam nitrazepam oxazepam temazepam	азепам	[no SUBSTEM]	клоназепам диазепам флуразепам лоразепам лорметазепам нитразепам оксазепам темазепам	azepam	[no SUBSTEM]	clonazepam diazepam flurazepam lorazepam lormetazepam nitrazepam oxazepam temazepam	diazepam derivatives
	azolam	alprazolam loprazolam midazolam		азолам	алпразолам лопразолам мидазолам		azolam	alprazolam loprazolam midazolam	
	variant: azam	clobazam		variant: азам	клобазам		variant: azam	clobazam	
	variant: opam	nefopam		variant: опам	нефопам		variant: opam	nefopam	
azocine	[no SUBSTEM]	pentazocine	азоцин	[no SUBSTEM]	пентазоцин	azocina	[no SUBSTEM]	pentazocina	narcotic antagonists/agonists related to 6,7-benzomorphan
azoline	[no SUBSTEM]	naphazoline oxymetazoline xylometazoline	азолин	[no SUBSTEM]	нафазолин оксиметазолин ксилометазолин	azolina	[no SUBSTEM]	nafazolina oximetazolina xilometazolina	antihistaminics or local vasoconstrictors, antazoline derivatives
azosin	[no SUBSTEM]	alfuzosin doxazosin prazosin tamsulosin terazosin	фузозин	[no SUBSTEM]	алфузозин доксазозин празозин тамсулозин теразозин	uzosina	[no SUBSTEM]	alfuzosina doxazosina prazosina tamsulosina terazosina	antihypertensive substances, prazosin derivatives
bactam	[no SUBSTEM]	tazobactam	бактам	[no SUBSTEM]	тазобактам	bactam	[no SUBSTEM]	tazobactam	β-lactamase inhibitors
bamate	[no SUBSTEM]	meprobamate	бамат	[no SUBSTEM]	мепробамат	bamato	[no SUBSTEM]	meprobamato	tranquillisers, propanediol and pentanediol derivatives

barb	[no SUBSTEM]	phenobarbital	барб	[no SUBSTEM]	фенобарбитал	barb	[no SUBSTEM]	fenobarbital	hypnotics, barbituric acid derivatives
begron	[no SUBSTEM]	mirabegron	бегрон	[no SUBSTEM]	мирабегрон	begron	[no SUBSTEM]	mirabegron	β3-adrenoreceptor agonists
bendazole	[no SUBSTEM]	mebendazole	бендазол	[no SUBSTEM]	мебендазол	bendazolo	[no SUBSTEM]	mebendazolo	anthelminthics, tiabendazole derivatives
bradine	[no SUBSTEM]	ivabradine	брадин	[no SUBSTEM]	ивабрадин	bradina	[no SUBSTEM]	ivabradina	bradycardic agents
bulin	[no SUBSTEM]	eribulin	булин	[no SUBSTEM]	эрибулин	bulina	[no SUBSTEM]	eribulina	antineoplastics; mitotic inhibitor, tubulin binder
cain	[no SUBSTEM]	flecainide	каин	[no SUBSTEM]	флекаинид	cain	[no SUBSTEM]	flecainide	class I antiarrhythmics, procainamide and lidocaine derivatives
caine	[no SUBSTEM]	benzocaine bupivacaine chloroprocaine cinchocaine levobupivacaine lidocaine oxybuprocaine pramocaine prilocaine procaine proxymetacaine ropivacaine tetracaine	каин	[no SUBSTEM]	бензокаин бупивакаин хлоропрокаин цинхокаин левобупивакаин лидокаин оксибупрокаин прамокаин прилокаин прокаин проксиметакаин ропивакаин тетракаин	caina	[no SUBSTEM]	benzocaina bupivacaina cloroprocaina cincocaina levobupivacaina lidocaina oxibuprocaina pramocaina prilocaina procaina prossimetacaina ropivacaina tetracaina	local anaesthetics
calci	[no SUBSTEM]	alfacalcidol calcipotriol calcitriol colecalciferol ergocalciferol	кальци	[no SUBSTEM]	альфакальцидол кальципотриол кальцитриол	calci	[no SUBSTEM]	alfacalcidolo calcipotriolo calcitriolo colecalciferolo ergocalciferolo	vitamin D analogues/derivatives

		paricalcitol tacalcitol			колекальциферол эргокальциферол парикальцитол такальцитол			paracalcitolo tacalcitolo	
capone	[no SUBSTEM]	entacapone opicapone tolcapone	капон	[no SUBSTEM]	энтакапон опикапон толкапон	capone	[no SUBSTEM]	entacapone opicapone tolcapone	catechol-O-methyltransferase (COMT) inhibitors
cef	[no SUBSTEM]	cefepime cefixime cefotaxime cefoxitin cefradine ceftazidime ceftobiprole ceftolozane ceftriaxone cefuroxime	цеф	[no SUBSTEM]	цефепим цефиксим цефотаксим цефокситин цефрадин цефтазидим цефтобипрол цефтолозан цефтриаксон цефуроксим	cef	[no SUBSTEM]	cefepima cefixima cefotaxima cefoxitina cefradina ceftazidima ceftobiprolo ceftolozane ceftriaxone cefuroxima	antibiotics, cefalosporanic acid derivatives
cel	[no SUBSTEM]	tisagenlecleucel	цел	[no SUBSTEM]	тисагенлеклейцел	cel	[no SUBSTEM]	tisagenlecleucel	cellulose derivatives
cept	bercept	aflibercept	цепт	берцепт	афлиберцепт	cept	bercept	aflibercept	receptor molecules or membrane ligands, native, modified or synthetic
	nercept	etanercept		нерцепт	этанерцепт		nercept	etanercept	
	tacept	abatacept belatacept		тацепт	абатацепт белатацепт		tacept	abatacept belatacept	
ciclib	[no SUBSTEM]	abemaciclib palbociclib ribociclib	циклиб	[no SUBSTEM]	абемациклиб	ciclib	[no SUBSTEM]	abemaciclib palbociclib ribociclib	cyclin dependant kinase inhibitors

					палбоциклиб рибоциклиб				
ciguat	[no SUBSTEM]	riociguat	цигуат	[no SUBSTEM]	риоцигуат	ciguat	[no SUBSTEM]	riociguat	guanylate cyclase activators and stimulators
cillin	[no SUBSTEM]	amoxicillin ampicillin benzylpenicillin flucloxacillin piperacillin	циллин	[no SUBSTEM]	амоксициллин ампициллин бензилпенициллин флуклоксациллин пиперациллин	cillina	[no SUBSTEM]	amoxicillina ampicillina benzilpenicillina flucloxacillina piperacillina	antibiotics, 6-aminopenicillanic acid derivatives
	cillinam	pivmecillinam		циллинам	пивмециллинам		cillinam	pivmecillinam	
clidine	[no SUBSTEM]	procyclidine	клидин	[no SUBSTEM]	проциклидин	clidina	[no SUBSTEM]	prociclidina	muscarinic receptor agonists/antagonists
clone	[no SUBSTEM]	zopiclone	клон	[no SUBSTEM]	зопиклон	clone	[no SUBSTEM]	zopiclone	hypnotic tranquillisers
conazole	[no SUBSTEM]	econazole fenticonazole fluconazole isavuconazole itraconazole ketoconazole miconazole posaconazole tioconazole voriconazole	коназол	[no SUBSTEM]	эконазол фентиконазол флуконазол изавуконазол итраконазол кетоконазол миконазол позаконазол тиоконазол вориконазол	conazolo	[no SUBSTEM]	econazolo fenticonazolo fluconazolo isavuconazolo itraconazolo ketoconazolo miconazolo posaconazolo tioconazolo voriconazolo	systemic antifungal agents, miconazole derivatives

cort	[no SUBSTEM]	deflazacort fludrocortisone fludroxycortide hydrocortisone	корт	[no SUBSTEM]	дефлазакорт флудрокортизон флудроксикортид гидрокортизон	cort	[no SUBSTEM]	deflazacort fludrocortisone fludroxicortide idrocortisone	corticosteroids, except prednisolone derivatives
coxib	[no SUBSTEM]	celecoxib etoricoxib parecoxib	коксиб	[no SUBSTEM]	целекоксиб эторикоксиб парекоксиб	coxib	[no SUBSTEM]	celecoxib etoricoxib parecoxib	selective cyclo-oxygenase inhibitors
cycline	[no SUBSTEM]	demeclocycline doxycycline lymecycline minocycline oxytetracycline tetracycline tigecycline	циклин	[no SUBSTEM]	демеклоциклин доксициклин лимециклин миноциклин окситетрациклин тетрациклин тигециклин	ciclina	[no SUBSTEM]	demeclociclina doxiciclina limeciclina minociclina oxitetraciclina tetraciclina tigeciclina	antibiotics, protein-synthesis inhibitors, tetracycline derivatives
dapsone	[no SUBSTEM]	dapsone	дапсон	[no SUBSTEM]	дапсон	dapsone	[no SUBSTEM]	dapsone	antimycobacterials, diaminodiphenylsulfone derivatives
degib	[no SUBSTEM]	vismodegib	дегиб	[no SUBSTEM]	висмодегиб	degib	[no SUBSTEM]	vismodegib	SMO receptor antagonists
dil	[no SUBSTEM]	minoxidil nicorandil	дил	[no SUBSTEM]	миноксидил никорандил	dil	[no SUBSTEM]	minoxidil nicorandil	vasodilators
	dilol	carvedilol		дилол	карведилол		dilolo	carvedilolo	
	dyl	bisacodyl trihexyphenidyl		дил	бисакодил тригексифенидил		dile	bisacodile triesifenidile	
dipine	[no SUBSTEM]	amlodipine felodipine	дипин	[no SUBSTEM]	амлодипин фелодипин	dipina	[no SUBSTEM]	amlodipina felodipina	calcium channel blockers, nifedipine derivatives

		lacidipine nifedipine nimodipine			лацидипин нифедипин нимодипин			lacidipina nifedipina nimodipina	
	nidipine	lercanidipine		нидипин	лерканидипин		nidipina	lercanidipina	
domide	[no SUBSTEM]	lenalidomide pomalidomide thalidomide	домид	[no SUBSTEM]	леналидомид помалидомид талидомид	domide	[no SUBSTEM]	lenalidomide pomalidomide talidomide	antineoplastics, thalidomide derivatives
dopa	[no SUBSTEM]	carbidopa levodopa methyldopa	допа	[no SUBSTEM]	карбидопа леводопа метилдопа	dopa	[no SUBSTEM]	carbidopa levodopa metildopa	dopamine receptor agonists, dopamine derivatives, used as antiparkinsonism/prolactin inhibitors
drine	[no SUBSTEM]	midodrine pseudoephedrine	дрин	[no SUBSTEM]	мидодрин псевдоэфедрин	drina	[no SUBSTEM]	midodrina pseudoefedrina	sympathomimetics
	frine	phenylephrine		фрин	фенилэфрин		frina	fenilefrina	
ectin	[no SUBSTEM]	ivermectin	ектин	[no SUBSTEM]	ивермектин	ectina	[no SUBSTEM]	ivermectina	antiparasitics, ivermectin derivatives
entan	[no SUBSTEM]	ambrisentan bosentan macitentan	ентан	[no SUBSTEM]	амбризентан бозентан мацитентан	entan	[no SUBSTEM]	ambrisentan bosentan macitentan	endothelin receptor antagonists
erg	[no SUBSTEM]	cabergoline ergometrine ergotamine	эрг	[no SUBSTEM]	каберголин эргометрин эрготамин	erg	[no SUBSTEM]	cabergolina ergometrina ergotamina	ergot alkaloid derivatives
ermin	sermin	mecasermin	ермин	зермин	меказермин	ermina	sermina	mecasermina	growth factors
estr	[no SUBSTEM]	estradiol estramustine mestranol	эстр	[no SUBSTEM]	эстрадиол эстрамустин местранол	estr	[no SUBSTEM]	estradiolo estramustina mestranolo	estrogens

fenacin	[no SUBSTEM]	darifenacin solifenacin	фенацин	[no SUBSTEM]	дарифенацин солифенацин	fenacina	[no SUBSTEM]	darifenacina solifenacina	muscarinic receptor antagonists
fentanil	[no SUBSTEM]	alfentanil fentanyl remifentanil sufentanil	фентанил	[no SUBSTEM]	алфентанил фентанил ремифентанил суфентанил	fentanil	[no SUBSTEM]	alfentanil fentanil remifentanil sufentanil	opioid receptor agonists, analgesics, fentanyl derivatives
fiban	[no SUBSTEM]	tirofiban	фибан	[no SUBSTEM]	тирофибан	fiban	[no SUBSTEM]	tirofiban	fibrinogen receptor antagonists (glycoprotein IIb/IIIa receptor antagonists)
fibrate	[no SUBSTEM]	bezafibrate fenofibrate	фибрат	[no SUBSTEM]	безафибрат фенофибрат	fibrato	[no SUBSTEM]	bezafibrato fenofibrato	clofibrate derivatives, peroxisome proliferator activated receptor-α (PPAR-α) agonists
	fibrozil	gemfibrozil		фиброзил	гемфиброзил		fibrozil	gemfibrozil	
flurane	[no SUBSTEM]	desflurane isoflurane methoxyflurane sevoflurane	флуран	[no SUBSTEM]	десфлуран изофлуран метоксифлуран севофлуран	flurano	[no SUBSTEM]	desflurano isoflurano metossiflurano sevoflurano	halogenated compounds used as general inhalation anaesthetics
formin	[no SUBSTEM]	metformin	формин	[no SUBSTEM]	метформин	formin	[no SUBSTEM]	metformina	antihyperglycaemics, phenformin derivatives
fos	fosfamide	cyclophosphamide ifosfamide	фос	фосфамид	циклофосфамид ифосфамид	fos	fosfamide	ciclofosfamide ifosfamide	insecticides, anthelminthics, pesticides etc., phosphorous derivatives
fungin	[no SUBSTEM]	anidulafungin caspofungin micafungin	фунгин	[no SUBSTEM]	анидулафунгин каспофунгин микафунгин	fungina	[no SUBSTEM]	anidulafungina caspofungina micafungina	antifungal antibiotics

fylline	[no SUBSTEM]	pentoxifylline	филлин	[no SUBSTEM]	пентоксифиллин	fillina	[no SUBSTEM]	pentoxifillina	N-methylated xanthine derivatives
gab	[no SUBSTEM]	gabapentin pregabalin retigabine tiagabine vigabatrin	габ	[no SUBSTEM]	габапентин прегабалин ретигабин тиагабин вигабатрин	gab	[no SUBSTEM]	gabapentin pregabalin retigabina tiagabina vigabatrina	gabamimetic agents
gado	[no SUBSTEM]	gadobutrol gadoteridol	гадо	[no SUBSTEM]	гадобутрол гадотеридол	gado	[no SUBSTEM]	gadobutrolo gadoteridolo	diagnostic agents, gadolinium derivatives
gatran	[no SUBSTEM]	dabigatran	гатран	[no SUBSTEM]	дабигатран	gatran	[no SUBSTEM]	dabigatran	thrombin inhibitor, antithrombotic agent
gest	[no SUBSTEM]	desogestrel dienogest dydrogesterone etonogestrel gestodene levonorgestrel medroxyprogesterone megestrol nomegestrol norelgestromin norgestimate progesterone	гест	[no SUBSTEM]	дезогестрел диеногест дидрогестерон этоногестрел гестоден левоноргестрел медроксипрогестерон мегэстрол номегэстрол норэлгестромин норгестимат прогестерон	gest	[no SUBSTEM]	desogestrel dienogest didrogesterone etonogestrel gestodene levonorgestrel medrossiprogesterone megestrolo nomegestrolo norelgestromin norgestimato progesterone	steroids, progestogens
giline	[no SUBSTEM]	rasagiline selegiline	гилин	[no SUBSTEM]	разагилин селегилин	gilina	[no SUBSTEM]	rasagilina selegilina	monoamine oxydase (MAO)-inhibitors type B
gli	[no SUBSTEM]	gliclazide glimepiride repaglinide	гли	[no SUBSTEM]	гликлазид глимепирид репаглинид	gli	[no SUBSTEM]	gliclazide glimepiride repaglinide	antihyperglycaemics

	amide	glibenclamide		амид	глибенкламид		amide	glibenclamide	
	gliflozin	canagliflozin dapagliflozin empagliflozin ertugliflozin		глифлозин	канаглифлозин дапаглифлозин эмпаглифлозин эртуглифлозин		gliflozin	canagliflozin dapagliflozin empagliflozin ertugliflozin	
	gliptin	alogliptin linagliptin saxagliptin sitagliptin vildagliptin		глиптин	алоглиптин линаглиптин саксаглиптин ситаглиптин вилдаглиптин		gliptin	alogliptin linagliptin saxagliptin sitagliptin vildagliptin	
	glitazone	pioglitazone		глитазон	пиоглитазон		glitazone	pioglitazone	
golide	[no SUBSTEM]	quinagolide	голид	[no SUBSTEM]	хинаголид	golide	[no SUBSTEM]	quinagolide	dopamine receptor agonists, ergoline derivatives
grel	[no SUBSTEM]	anagrelide cangrelor clopidogrel prasugrel ticagrelor	грел	[no SUBSTEM]	анагрелид	grel	[no SUBSTEM]	anagrelide cangrelor clopidogrel prasugrel ticagrelor	platelet aggregation inhibitors
guan	[no SUBSTEM]	guanethidine	гуан	[no SUBSTEM]	гуанетидин	guan	[no SUBSTEM]	guanetidina	antihypertensives, guanidine derivatives
icam	[no SUBSTEM]	meloxicam piroxicam tenoxicam	икам	[no SUBSTEM]	мелоксикам пироксикам теноксикам	icam	[no SUBSTEM]	meloxicam piroxicam tenoxicam	anti-inflammatory, isoxicam derivatives
ifene	mifene	clomifene ospemifene toremifene	ифен	мифен	кломифен оспемифен торемифен	ifene	mifene	clomifene ospemifene toremifene	antiestrogens or estrogen receptor modulators,

	oxifene	bazedoxifene raloxifene tamoxifen		оксифен	базедоксифен ралоксифен тамоксифен		oxifene	bazedoxifene raloxifene tamoxifene	clomifene and tamoxifen derivatives
imibe	[no SUBSTEM]	ezetimibe	имиб	[no SUBSTEM]	эзетимиб	imibe	[no SUBSTEM]	ezetimibe	antihyperlipidaemics, acyl CoA: cholesterol acyltransferase (ACAT) inhibitors
imod	[no SUBSTEM]	fingolimod imiquimod	имод	[no SUBSTEM]	финголимод имихимод	imod	[no SUBSTEM]	fingolimod imiquimod	immunomodulators, both stimulant/suppressive and stimulant
imus	rolimus	everolimus pimecrolimus sirolimus tacrolimus temsirolimus	имус	ролимус	эверолимус пимекролимус сиролимус такролимус темсиролимус	imus	rolimus	everolimus pimecrolimus sirolimus tacrolimus temsirolimus	immunosuppressants (other than antineoplastics)
io	[no SUBSTEM]	iomeprol iopamidol	йо	[no SUBSTEM]	йомепрол йопамидол	io	[no SUBSTEM]	iomeprolo iopamidolo	iodine-containing contrast media
irudin	[no SUBSTEM]	bivalirudin	ирудин	[no SUBSTEM]	бивалирудин	irudina	[no SUBSTEM]	bivalirudina	thrombin inhibitors, hirudin derivatives
isant	[no SUBSTEM]	pitolisant	изант	[no SUBSTEM]	питолизант	isant	[no SUBSTEM]	pitolisant	histamine H3 receptor antagonists
izine	[no SUBSTEM]	buclizine cetirizine cyclizine hydroxyzine levocetirizine	изин	[no SUBSTEM]	буклизин цетиризин циклизин гидроксизин левоцетиризин	izina	[no SUBSTEM]	buclizina cetirizina ciclizina idroxizina levocetirizina	diphenylmethyl piperazine derivatives
	rizine	cinnarizine		ризин	циннаризин		rizina	cinnarizina	
kacin	[no SUBSTEM]	amikacin	кацин	[no SUBSTEM]	амикацин	kacina	[no SUBSTEM]	amikacina	antibiotics, kanamycin and bekanamycin derivatives

									(obtained from Streptomyces kanamyceticus)
kalant	[no SUBSTEM]	vernakalant	калант	[no SUBSTEM]	вернакалант	kalant	[no SUBSTEM]	vernakalant	potassium channel blockers
kin	leukin	aldesleukin	кин	лейкин	алдеслейкин	chin	leuchina	aldesleuchina	interleukin type substances
kinra	nakinra	anakinra	кинра	накинра	анакинра	kinra	nakinra	anakinra	interleukin receptor antagonists
lisib	[no SUBSTEM]	idelalisib	лисиб	[no SUBSTEM]	иделалисиб	lisib	[no SUBSTEM]	idelalisib	phosphatidylinositol 3-kinase inhibitors, antineoplastics
lutamide	[no SUBSTEM]	apalutamide bicalutamide enzalutamide flutamide	лутамид	[no SUBSTEM]	апалутамид бикалутамид энзалутамид флутамид	lutamide	[no SUBSTEM]	apalutamide bicalutamide enzalutamide flutamide	non-steroid antiandrogens
mab	limab	cemiplimab	маб	лимаб	цемиплимаб	mab	limab	cemiplimab	monoclonal antibodies
	omab	blinatumomab		омаб	блинатумомаб		omab	blinatumomab	
	umab	adalimumab alirocumab avelumab belimumab bezlotoxumab brodalumab burosumab canakinumab daratumumab denosumab dupilumab durvalumab erenumab evolocumab		умаб	адалимумаб алирокумаб авелумаб белимумаб безлотоксумаб бродалумаб буросумаб канакинумаб даратумумаб деносумаб дупилумаб дурвалумаб		umab	adalimumab alirocumab avelumab belimumab bezlotoxumab brodalumab burosumab canakinumab daratumumab denosumab dupilumab durvalumab erenumab evolocumab	

		golimumab guselkumab ipilimumab lanadelumab necitumumab nivolumab olaratumab panitumumab ramucirumab sarilumab secukinumab ustekinumab			эренумаб эволокумаб голимумаб гуселкумаб ипилимумаб ланаделумаб нецитумумаб ниволумаб оларатумаб панитумумаб рамуцирумаб сарилумаб секукинумаб устекинумаб			golimumab guselkumab ipilimumab lanadelumab necitumumab nivolumab olaratumab panitumumab ramucirumab sarilumab secukinumab ustekinumab	
	ximab	basiliximab cetuximab dinutuximab infliximab rituximab siltuximab		ксимаб	базиликсимаб цетуксимаб динутуксимаб инфликсимаб ритуксимаб силтуксимаб		ximab	basiliximab cetuximab dinutuximab infliximab rituximab siltuximab	
	zumab	alemtuzumab atezolizumab benralizumab bevacizumab		зумаб	алемтузумаб атезолизумаб бенрализумаб бевацизумаб		zumab	alemtuzumab atezolizumab benralizumab bevacizumab	

		caplacizumab eculizumab elotuzumab emicizumab fremanezumab galcanezumab gemtuzumab idarucizumab ixekizumab mepolizumab natalizumab obinutuzumab ocrelizumab omalizumab palivizumab pembrolizumab pertuzumab ranibizumab ravulizumab reslizumab risankizumab tildrakizumab tocilizumab trastuzumab vedolizumab			каплацизумаб экулизумаб элотузумаб эмицизумаб фреманезумаб галканезумаб гемтузумаб идаруцизумаб иксекизумаб меполизумаб натализумаб обинутузумаб окрелизумаб омализумаб паливизумаб пембролизумаб пертузумаб ранибизумаб равулизумаб реслизумаб рисанкизумаб тилдракизумаб тоцилизумаб			caplacizumab eculizumab elotuzumab emicizumab fremanezumab galcanezumab gemtuzumab idarucizumab ixekizumab mepolizumab natalizumab obinutuzumab ocrelizumab omalizumab palivizumab pembrolizumab pertuzumab ranibizumab ravulizumab reslizumab risankizumab tildrakizumab tocilizumab trastuzumab vedolizumab	

					трастузумаб ведолизумаб				
mantine	[no SUBSTEM]	memantine	мантин	[no SUBSTEM]	мемантин	mantina	[no SUBSTEM]	memantina	adamantane derivatives
mer	[no SUBSTEM]	carbomer sevelamer	мер	[no SUBSTEM]	карбомер севеламер	mer	[no SUBSTEM]	carbomero sevelamer	mercury-containing drugs, antimicrobial or diuretic (deleted from General Principles in List 28 prop. INN)
mestane	[no SUBSTEM]	exemestane	местан	[no SUBSTEM]	эксеместан	mestane	[no SUBSTEM]	exemestane	aromatase inhibitors
metacin	[no SUBSTEM]	acemetacin indometacin	метацин	[no SUBSTEM]	ацеметацин индометацин	metacina	[no SUBSTEM]	acemetacina indometacina	anti-inflammatory, indometacin derivatives
micin	[no SUBSTEM]	fidaxomicin gentamicin ozogamicin	мицин	[no SUBSTEM]	фидаксомицин гентамицин озогамицин	micina	[no SUBSTEM]	fidaxomicina gentamicina ozogamicina	aminoglycosides, antibiotics obtained from various Micromonospora
mito	[no SUBSTEM]	mitotane	мито	[no SUBSTEM]	митотан	mito	[no SUBSTEM]	mitotane	antineoplastics, nucleotoxic agents (deleted from General Principles in List 24 prop. INN)
mustine	[no SUBSTEM]	bendamustine carmustine lomustine	мустин	[no SUBSTEM]	бендамустин кармустин ломустин	mustina	[no SUBSTEM]	bendamustina carmustina lomustina	antineoplastic, alkylating agents, (β-chloroethyl)
mycin	[no SUBSTEM]	azithromycin capreomycin clarithromycin clindamycin dactinomycin	мицин	[no SUBSTEM]	азитромицин капреомицин кларитромицин клиндамицин	micina	[no SUBSTEM]	azitromicina capreomicina claritromicina clindamicina dactinomicina	antibiotics, produced by Streptomyces strains (see also -kacin)

		daptomycin erythromycin fosfomycin mitomycin neomycin vancomycin			дактиномицин даптомицин эритромицин фосфомицин митомицин неомицин ванкомицин			daptomicina eritromicina fosfomicina mitomicina neomicina vancomicina	
nab	[no SUBSTEM]	cannabidiol	наб	[no SUBSTEM]	каннабидиол	nab	[no SUBSTEM]	cannabidiol	cannabinoid receptors agonists
nal	[no SUBSTEM]	nalmefene naloxegol naloxone naltrexone	нал	[no SUBSTEM]	налмефен налоксегол налоксон налтрексон	nal	[no SUBSTEM]	nalmefene naloxegol naloxone naltrexone	opioid receptor antagonists/agonists related to normorphine
nicline	[no SUBSTEM]	varenicline	никлин	[no SUBSTEM]	вареniклин	niclina	[no SUBSTEM]	vareniclina	nicotinic acetylcholine receptor partial agonists / agonists
nic	[no SUBSTEM]	nicardipine nicotinamide	ник	[no SUBSTEM]	никардипин никотинамид	nic	[no SUBSTEM]	nicardipina nicotinamide	nicotinic acid or nicotinoyl alcohol derivatives
nidazole	[no SUBSTEM]	metronidazole	нидазол	[no SUBSTEM]	метронидазол	nidazolo	[no SUBSTEM]	metronidazolo	antiprotozoals and radiosensitisers, metronidazole derivatives
nitro	[no SUBSTEM]	nitrofurantoin	нитро	[no SUBSTEM]	нитрофурантоин	nitro	[no SUBSTEM]	nitrofurantoina	NO2 - derivatives
olol	[no SUBSTEM]	acebutolol atenolol betaxolol bisoprolol celiprolol esmolol levobunolol metoprolol nadolol nebivolol	олол	[no SUBSTEM]	ацебутолол атенолол бетаксолол бисопролол целипролол эсмолол левобуналол метопролол надолол	ololo	[no SUBSTEM]	acebutololo atenololo betaxololo bisoprololo celiprololo esmololo levobunololo metoprololo nadololo nebivololo	β-adrenoreceptor antagonists

		oxprenolol pindolol propranolol timolol			небиволол окспреноло л пиндолол пропраноло л тимолол			oxprenololo pindololo propranololo timololo	
	alol	labetalol sotalol		алол	лабеталол соталол		alolo	labetalolo sotalolo	
onide	[no SUBSTEM]	budesonide ciclesonide	онид	[no SUBSTEM]	будесонид циклесонид	onide	[no SUBSTEM]	budesonide ciclesonide	steroids for topical use, acetal derivatives
onidine	[no SUBSTEM]	apraclonidin e brimonidine clonidine moxonidine	онидин	[no SUBSTEM]	апраклонид ин бримониди н клонидин моксонидин	onidina	[no SUBSTEM]	apraclonidin a brimonidina clonidina moxonidina	antihypertensives, clonidine derivatives
orex	[no SUBSTEM]	bupropion	(exocentric name, transliterate d in full)	[no SUBSTEM]	бупропион	(exocentric name)	[no SUBSTEM]	bupropione	anorexics
orphan	orphine	buprenorphi ne	орфан	орфин	бупренорф ин	orfano	orfina	buprenorfina	opioid receptor antagonists/agonists, morphinan derivates
	orphone	hydromorph one		орфон	гидроморф он		orfone	idromorfone	
oxacin	[no SUBSTEM]	ciprofloxacin levofloxacin moxifloxacin ofloxacin	оксацин	[no SUBSTEM]	ципрофлок сацин левофлокс ацин моксифлокс ацин офлоксаци н	oxacina	[no SUBSTEM]	ciprofloxacin a levofloxacin a moxifloxacin a ofloxacina	antibacterials, nalidixic acid derivatives
oxetine	[no SUBSTEM]	atomoxetine dapoxetine duloxetine fluoxetine	оксетин	[no SUBSTEM]	атомоксети н дапоксетин дулоксетин	oxetina	[no SUBSTEM]	atomoxetina dapoxetina duloxetina fluoxetina	serotonin and/or norepinephrine reuptake inhibitors, fluoxetine derivatives

		paroxetine reboxetine vortioxetine			флуоксетин пароксетин ребоксетин вортиоксетин			paroxetina reboxetina vortioxetine	
pamide	[no SUBSTEM]	indapamide xipamide	памид	[no SUBSTEM]	индапамид ксипамид	pamide	[no SUBSTEM]	indapamide xipamide	diuretics, sulfamoylbenzoic acid derivatives (could be sulfamoylbenzamide)
pamil	[no SUBSTEM]	verapamil	памил	[no SUBSTEM]	верапамил	pamil	[no SUBSTEM]	verapamil	calcium channel blocker, verapamil derivatives
parib	[no SUBSTEM]	niraparib olaparib rucaparib	париб	[no SUBSTEM]	нирапариб олапариб рукапариб	parib	[no SUBSTEM]	niraparib olaparib rucaparib	poly-ADP-Ribose polymerase inhibitors
penem	[no SUBSTEM]	ertapenem imipenem meropenem	пенем	[no SUBSTEM]	эртапенем имипенем меропенем	penem	[no SUBSTEM]	ertapenem imipenem meropenem	analogues of penicillanic acid antibiotics modified in the five-membered ring
perfl	[no SUBSTEM]	perflutren	перфл	[no SUBSTEM]	перфлутрен	perfl	[no SUBSTEM]	perflutreno	perfluorinated compounds used as blood
peridol	[no SUBSTEM]	benperidol droperidol haloperidol	перидол	[no SUBSTEM]	бенперидол дроперидол галоперидол	peridolo	[no SUBSTEM]	benperidolo droperidolo aloperidolo	substitutes and/or diagnostic agents
peridone	[no SUBSTEM]	domperidone paliperidone risperidone	перидон	[no SUBSTEM]	домперидон палиперидон рисперидон	peridone	[no SUBSTEM]	domperidone paliperidone risperidone	antipsychotics, risperidone derivatives
pidem	[no SUBSTEM]	zolpidem	пидем	[no SUBSTEM]	золпидем	pidem	[no SUBSTEM]	zolpidem	hypnotics/sedatives, zolpidem derivatives
pine	apine	asenapine clozapine loxapine mirtazapine olanzapine quetiapine	пин	апин	азенапин клозапин локсапин миртазапин оланзапин кветиапин	pina	apina	asenapina clozapina loxapina mirtazapina olanzapina quetiapina	tricyclic compounds

	oxepin	doxepin		оксепин	доксепин		oxepina	doxepina	
	zepine	carbamazepine eslicarbazepine oxcarbazepine		зепин	карбамазепин эсликарбазепин окскарбазепин		zepina	carbamazepina eslicarbazepina oxcarbazepina	
planin	[no SUBSTEM]	teicoplanin	планин	[no SUBSTEM]	тейкопланин	planina	[no SUBSTEM]	teicoplanina	glycopeptide antibacterials (Actinoplanes strains)
platin	[no SUBSTEM]	carboplatin cisplatin oxaliplatin	платин	[no SUBSTEM]	карбоплатин цисплатин оксалиплатин	platino	[no SUBSTEM]	carboplatino cisplatino oxaliplatino	antineoplastic agents, platinum derivatives
porfin	[no SUBSTEM]	padeliporfin verteporfin	порфин	[no SUBSTEM]	паделипорфин вертепорфин	porfina	[no SUBSTEM]	padeliporfina verteporfina	benzoporphyrin derivatives
pramine	[no SUBSTEM]	clomipramine imipramine lofepramine trimipramine	прамин	[no SUBSTEM]	кломипрамин имипрамин лофепрамин тримипрамин	pramina	[no SUBSTEM]	clomipramina imipramina lofepramina trimipramina	substances of the imipramine group
prazole	[no SUBSTEM]	aripiprazole esomeprazole lansoprazole omeprazole pantoprazole rabeprazole	празол	[no SUBSTEM]	арипипразол эзомепразол лансопразол омепразол пантопразол рабепразол	prazolo	[no SUBSTEM]	aripiprazolo esomeprazolo lansoprazolo omeprazolo pantoprazolo rabeprazolo	antiulcer, benzimidazole derivatives
pred	[no SUBSTEM]	loteprednol	пред	[no SUBSTEM]	лотепреднол	pred	[no SUBSTEM]	loteprednol	prednisone and prednisolone derivatives

		methylprednisolone prednisolone			метилпреднизолон преднизолон			metilprednisolone prednisolone	
	olone	tibolone		олон	тиболон		olone	tibolone	
pressin	[no SUBSTEM]	argipressin desmopressin terlipressin	прессин	[no SUBSTEM]	аргипрессин десмопрессин терлипрессин	pressina	[no SUBSTEM]	argipressina desmopressina terlipressina	vasoconstrictors, vasopressin derivatives
pride	[no SUBSTEM]	amisulpride prucalopride	прид	[no SUBSTEM]	амисульприд прукалоприд	pride	[no SUBSTEM]	amisulpride prucalopride	sulpiride derivatives
pril	[no SUBSTEM]	captopril enalapril fosinopril imidapril lisinopril moexipril perindopril quinapril ramipril trandolapril	прил	[no SUBSTEM]	каптоприл эналаприл фозиноприл имидаприл лизиноприл моэксиприл периндоприл хинаприл рамиприл трандолаприл	pril	[no SUBSTEM]	captopril enalapril fosinopril imidapril lisinopril moexipril perindopril quinapril ramipril trandolapril	angiotensin-converting enzyme inhibitors
prim	[no SUBSTEM]	trimethoprim	прим	[no SUBSTEM]	триметоприм	prim	[no SUBSTEM]	trimetoprim	antibacterials, dihydrofolate reductase (DHFR) inhibitors, trimethoprim derivatives
pris	[no SUBSTEM]	mifepristone ulipristal	прис	[no SUBSTEM]	мифепристон улипристал	pris	[no SUBSTEM]	mifepristone ulipristal	steroidal compounds acting on progesterone receptors (excluding -gest- compounds)
profen	[no SUBSTEM]	dexketoprofen flurbiprofen	профен	[no SUBSTEM]	декскетопрофен	profene	[no SUBSTEM]	dexketoprofene flurbiprofene	anti-inflammatory agents, ibuprofen derivatives

		ibuprofen ketoprofen			флурбипрофен ибупрофен кетопрофен			ibuprofene ketoprofene	
prost	[no SUBSTEM]	alprostadil bimatoprost dinoprostone epoprostenol gemeprost iloprost latanoprost misoprostol tafluprost travoprost	прост	[no SUBSTEM]	алпростадил биматопрост динопростон эпопростенол гемепрост илопрост латанопрост мизопростол тафлупрост травопрост	prost	[no SUBSTEM]	alprostadil bimatoprost dinoprostone epoprostenolo gemeprost iloprost latanoprost misoprostolo tafluprost travoprost	prostaglandins
quine	[no SUBSTEM]	chloroquine hydroxychloroquine mefloquine	хин	[no SUBSTEM]	хлорохин гидроксихлорохин мефлохин	china	[no SUBSTEM]	clorochina idroxiclorochina meflochina	quinoline derivatives (deleted from General Principles in List 28 prop. INN)
racetam	[no SUBSTEM]	brivaracetam levetiracetam piracetam	рацетам	[no SUBSTEM]	бриварацетам леветирацетам пирацетам	racetam	[no SUBSTEM]	brivaracetam levetiracetam piracetam	amide type nootrope agents, piracetam derivatives
racil	[no SUBSTEM]	fluorouracil gimeracil oteracil	рацил	[no SUBSTEM]	фторурацил гимерацил отерацил	racile	[no SUBSTEM]	fluorouracile gimeracile oteracile	uracil type antineoplastics
	thiouracil	propylthiouracil		тиоурацил	пропилтиоурацил		tiouracile	propiltiouracile	
rafenib	[no SUBSTEM]	dabrafenib encorafenib regorafenib	рафениб	[no SUBSTEM]	дабрафениб	rafenib	[no SUBSTEM]	dabrafenib encorafenib regorafenib	Raf (rapidly accelerated fibrosarcoma) kinase inhibitors

		sorafenib vemurafenib			энкорафениб регорафениб сорафениб вемурафениб			sorafenib vemurafenib	
relin	[no SUBSTEM]	buserelin gonadorelin goserelin leuprorelin nafarelin somatorelin triptorelin	релин	[no SUBSTEM]	бусерелин гонадорелин гозерелин лейпрорелин нафарелин соматорелин трипторелин	relina	[no SUBSTEM]	buserelina gonadorelina goserelina leuprorelina nafarelina somatorelina triptorelina	pituitary hormone-release stimulating peptides
relix	[no SUBSTEM]	cetrorelix degarelix ganirelix	реликс	[no SUBSTEM]	цетрореликс дегареликс ганиреликс	relix	[no SUBSTEM]	cetrorelix degarelix ganirelix	gonadotropin-releasing-hormone (GnRH) inhibitors, peptides
renone	[no SUBSTEM]	drospirenone eplerenone	ренон	[no SUBSTEM]	дроспиренон эплеренон	renone	[no SUBSTEM]	drospirenone eplerenone	aldosterone antagonists, spironolactone derivates
retin	[no SUBSTEM]	acitretin alitretinoin isotretinoin retinol tretinoin	ретин	[no SUBSTEM]	ацитретин алитретиноин изотретиноин ретинол третиноин	retin	[no SUBSTEM]	acitretina alitretinoina isotretinoina retinolo tretinoina	retinol derivatives
ribine	[no SUBSTEM]	cladribine	рибин	[no SUBSTEM]	кладрибин	ribine	[no SUBSTEM]	cladribine	ribofuranyl-derivatives of the “pyrazofurin” type
rifa	[no SUBSTEM]	rifabutin rifampicin rifaximin	рифа	[no SUBSTEM]	рифабутин рифампицин	rifa	[no SUBSTEM]	rifabutina rifampicina rifaximina	antibiotics, rifamycin derivatives

					рифаксимин				
rozole	[no SUBSTEM]	anastrozole letrozole	розол	[no SUBSTEM]	анастрозол летрозол	rozolo	[no SUBSTEM]	anastrozolo letrozolo	aromatase inhibitors, imidazole-triazole derivatives
rsen	[no SUBSTEM]	inotersen nusinersen	рсен	[no SUBSTEM]	инотерсен нусинерсен	rsen	[no SUBSTEM]	inotersen nusinersen	antisense oligonucleotides
rubicin	[no SUBSTEM]	daunorubicin doxorubicin epirubicin idarubicin	рубицин	[no SUBSTEM]	даунорубицин доксорубицин эпирубицин идарубицин	rubicina	[no SUBSTEM]	daunorubicina doxorubicina epirubicina idarubicina	antineoplastics, daunorubicin derivatives
sal	salazide	balsalazide	сал	салазид	балсалазид	sal	salazide	balsalazide	salicylic acid derivatives
	salazine	mesalazine olsalazine sulfasalazine		салазин	месалазин олсалазин сульфасалазин		salazina	mesalazina olsalazina sulfasalazina	salicylic acid derivatives
sartan	[no SUBSTEM]	azilsartan candesartan eprosartan irbesartan losartan olmesartan telmisartan valsartan	сартан	[no SUBSTEM]	азилсартан кандесартан эпросартан ирбесартан лозартан олмесартан телмисартан валсартан	sartan	[no SUBSTEM]	azilsartan candesartan eprosartan irbesartan losartan olmesartan telmisartan valsartan	angiotensin II receptor antagonists, antihypertensive (non-peptidic)
semide	[no SUBSTEM]	furosemide torasemide	семид	[no SUBSTEM]	фуросемид торасемид	semide	[no SUBSTEM]	furosemide torasemide	diuretics, furosemide derivatives
setron	[no SUBSTEM]	granisetron ondansetron palonosetron	сетрон	[no SUBSTEM]	гранисетрон ондансетрон палоносетрон	setron	[no SUBSTEM]	granisetron ondansetron palonosetron	oxytocin antagonists

siran	[no SUBSTEM]	patisiran	сиран	[no SUBSTEM]	патисиран	siran	[no SUBSTEM]	patisiran	small interfering RNA
som	[no SUBSTEM]	somatropin	сом	[no SUBSTEM]	соматропин	som	[no SUBSTEM]	somatropina	growth hormone derivatives
spirone	[no SUBSTEM]	buspirone	спирон	[no SUBSTEM]	буспирон	spirone	[no SUBSTEM]	buspirone	anxiolytics, buspirone derivatives
stat	[no SUBSTEM]	cobicistat eliglustat febuxostat miglustat pentostatin telotristat	стат	[no SUBSTEM]	кобицистат элиглустат фебуксостат миглустат пентостатин телотристат	stat	[no SUBSTEM]	cobicistat eliglustat febuxostat miglustat pentostatina telotristat	enzyme inhibitors
	inostat	panobinostat		иностат	панобиностат		inostat	panobinostat	
	listat	orlistat		листат	орлистат		listat	orlistat	
	vastatin	atorvastatin fluvastatin pravastatin rosuvastatin simvastatin		вастатин	аторвастатин флувастатин правастатин розувастатин симвастатин		vastatina	atorvastatina fluvastatina pravastatina rosuvastatina simvastatina	
steine	[no SUBSTEM]	acetylcysteine carbocisteine erdosteine	стеин	[no SUBSTEM]	ацетилцистеин карбоцистеин эрдостеин	steina	[no SUBSTEM]	acetilcisteina carbocisteina erdosteina	mucolytics, other than bromhexine derivatives
ster	sterol	dihydrotachysterol	стер	стерол	дигидротахистерол	ster	sterolo	diidrotachisterolo	androgens/anabolic steroids

steride	[no SUBSTEM]	dutasteride finasteride	стерид	[no SUBSTEM]	дутастерид финастерид	steride	[no SUBSTEM]	dutasteride finasteride	androgens/anabolic steroids
stigmine	[no SUBSTEM]	rivastigmine	стигмин	[no SUBSTEM]	ривастигмин	stigmina	[no SUBSTEM]	rivastigmina	acetylcholinesterase inhibitors
stim	[no SUBSTEM]	romiplostim	стим	[no SUBSTEM]	ромиплостим	stim	[no SUBSTEM]	romiplostim	colony stimulating factors
	grastim	filgrastim lenograstim lipegfilgrastim pegfilgrastim		грастим	филграстим ленограстим липэгфилграстим пэгфилграстим		grastim	filgrastim lenograstim lipegfilgrastim pegfilgrastim	
sulfa	[no SUBSTEM]	sulfadiazine sulfamethoxazole	сульфа	[no SUBSTEM]	сульфадиазин сульфаметоксазол	sulfa	[no SUBSTEM]	sulfadiazina sulfametoxazolo	anti-infectives, sulfonamides
sulfan	[no SUBSTEM]	busulfan treosulfan	сульфан	[no SUBSTEM]	бусульфан треосульфан	sulfano	[no SUBSTEM]	busulfano treosulfano	antineoplastic, alkylating agents, methanesulfonates
tadine	[no SUBSTEM]	desloratadine loratadine olopatadine rupatadine	тадин	[no SUBSTEM]	дезлоратадин лоратадин олопатадин рупатадин	tadina	[no SUBSTEM]	desloratadina loratadina olopatadina rupatadina	tricyclic histamine-H1 receptor antagonists, tricyclic compounds
tant	pitant	aprepitant fosaprepitant netupitant rolapitant	тант	питант	апрепитант фосапрепитант нетупитант ролапитант	tant	pitant	aprepitant fosaprepitant netupitant rolapitant	neurokinin (tachykinin) receptor antagonists
tapide	[no SUBSTEM]	lomitapide	тапид	[no SUBSTEM]	ломитапид	tapide	[no SUBSTEM]	lomitapide	microsomal triglyceride transfer protein (MTP) inhibitors
taxel	[no SUBSTEM]	cabazitaxel docetaxel	таксел	[no SUBSTEM]	кабазитаксел	taxel	[no SUBSTEM]	cabazitaxel docetaxel	antineoplastics; taxane derivatives

		paclitaxel			доцетаксел паклитаксел			paclitaxel	
tecan	[no SUBSTEM]	irinotecan	текан	[no SUBSTEM]	иринотекан	tecan	[no SUBSTEM]	irinotecan	antineoplastics, topoisomerase I inhibitors
terol	[no SUBSTEM]	formoterol indacaterol olodaterol salmeterol vilanterol	терол	[no SUBSTEM]	формотерол индакатерол олодатерол салметерол вилантерол	terolo	[no SUBSTEM]	formoterolo indacaterolo olodaterolo salmeterolo vilanterolo	bronchodilators, phenethylamine derivatives
	buterol	bambuterol		бутерол	бамбутерол		buterolo	bambuterolo	
terone	[no SUBSTEM]	abiraterone cyproterone	терон	[no SUBSTEM]	абиратерон ципротерон	terone	[no SUBSTEM]	abiraterone ciproterone	antiandrogens
tiazem	[no SUBSTEM]	diltiazem	тиазем	[no SUBSTEM]	дилтиазем	tiazem	[no SUBSTEM]	diltiazem	calcium channel blockers, diltiazem derivatives
tibant	[no SUBSTEM]	icatibant	тибант	[no SUBSTEM]	икатибант	tibant	[no SUBSTEM]	icatibant	bradykinin receptor antagonists
tide	[no SUBSTEM]	eptifibatide etelcalcetide exenatide linaclotide lixisenatide mifamurtide teriparatide ziconotide	тид	[no SUBSTEM]	эптифибатид этелкальцетид эксенатид линаклотид ликсисенатид мифамуртид терипаратид зиконотид	tide	[no SUBSTEM]	eptifibatide etelcalcetide exenatide linaclotide lixisenatide mifamurtide teriparatide ziconotide	peptides and glycopeptides (for special groups of peptides see -actide, -pressin, -relin, -tocin)
	glutide	albiglutide dulaglutide liraglutide semaglutide		глутид	албиглутид дулаглутид лираглутид семаглутид		glutide	albiglutide dulaglutide liraglutide semaglutide	

		teduglutide			тедуглутид			teduglutide	
	reotide	lanreotide octreotide pasireotide		реотид	ланреотид октреотид пасиреотид		reotide	lanreotide octreotide pasireotide	
tidine	[no SUBSTEM]	cimetidine nizatidine ranitidine	тидин	[no SUBSTEM]	циметидин низатидин ранитидин	tidina	[no SUBSTEM]	cimetidina nizatidina ranitidina	histamine-H2-receptor antagonists, cimetidine derivatives
tinib	[no SUBSTEM]	afatinib alectinib axitinib bosutinib brigatinib cabozantinib ceritinib crizotinib dacomitinib dasatinib erlotinib gefitinib imatinib lapatinib larotrectinib lenvatinib lorlatinib nilotinib osimertinib ponatinib ruxolitinib sunitinib	тиниб	[no SUBSTEM]	афатиниб алектиниб акситиниб бозутиниб бригатиниб кабозантиниб церитиниб кризотиниб дакомитиниб дазатиниб эрлотиниб гефитиниб иматиниб лапатиниб ларотректиниб ленватиниб лорлатиниб нилотиниб осимертиниб понатиниб руксолитиниб сунитиниб	tinib	[no SUBSTEM]	afatinib alectinib axitinib bosutinib brigatinib cabozantinib ceritinib crizotinib dacomitinib dasatinib erlotinib gefitinib imatinib lapatinib larotrectinib lenvatinib lorlatinib nilotinib osimertinib ponatinib ruxolitinib sunitinib	tyrosine kinase inhibitors
	brutinib	ibrutinib		брутиниб	ибрутиниб		brutinib	ibrutinib	
	citinib	baricitinib tofacitinib		цитиниб	барицитиниб		citinib	baricitinib tofacitinib	

		upadacitinib			тофацитиниб упадацитиниб			upadacitinib	
	metinib	binimetinib cobimetinib trametinib		метиниб	биниметиниб кобиметиниб траметиниб		metinib	binimetinib cobimetinib trametinib	
tocin	[no SUBSTEM]	carbetocin oxytocin	тоцин	[no SUBSTEM]	карбетоцин окситоцин	tocina	[no SUBSTEM]	carbetocina oxitocina	oxytocin derivatives
toin	[no SUBSTEM]	fosphenytoin phenytoin	тоин	[no SUBSTEM]	фосфенитоин фенитоин	toina	[no SUBSTEM]	fosfenitoina fenitoina	antiepileptics, hydantoin derivatives
traline	[no SUBSTEM]	sertraline	тралин	[no SUBSTEM]	сертралин	tralina	[no SUBSTEM]	sertralina	serotonin reuptake inhibitors
trexate	[no SUBSTEM]	methotrexate	трексат	[no SUBSTEM]	метотрексат	trexato	[no SUBSTEM]	metotrexato	folic acid analogues
trexed	[no SUBSTEM]	pemetrexed raltitrexed	трексед	[no SUBSTEM]	пеметрексед ралтитрексед	trexed	[no SUBSTEM]	pemetrexed raltitrexed	antineoplastics; thymidilate synthetase inhibitors
trigine	[no SUBSTEM]	lamotrigine	триджин	[no SUBSTEM]	ламотриджин	trigina	[no SUBSTEM]	lamotrigina	sodium channel blockers, signal transduction modulators
tril	[no SUBSTEM]	sacubitril	трил	[no SUBSTEM]	сакубитрил	tril	[no SUBSTEM]	sacubitril	endopeptidase inhibitors
	dotril	racecadotril		дотрил	рацекадотрил		dotril	racecadotril	
triptan	[no SUBSTEM]	almotriptan eletriptan frovatriptan rizatriptan sumatriptan zolmitriptan	триптан	[no SUBSTEM]	алмотриптан элетриптан фроватриптан	triptan	[no SUBSTEM]	almotriptan eletriptan frovatriptan rizatriptan sumatriptan zolmitriptan	serotonin (5HT1) receptor agonists, sumatriptan derivatives

					ризатриптан суматриптан золмитриптан				
triptyline	[no SUBSTEM]	amitriptyline nortriptyline	триптилин	[no SUBSTEM]	амитриптилин нортриптилин	triptilina	[no SUBSTEM]	amitriptilina nortriptilina	antidepressants, dibenzo[a,d]cycloheptane or cyclopheptene derivatives
troban	[no SUBSTEM]	argatroban	тробан	[no SUBSTEM]	аргатробан	troban	[no SUBSTEM]	argatroban	thromboxane A2-receptor antagonists; antithrombotic agents
trop	[no SUBSTEM]	tropicamide	троп	[no SUBSTEM]	тропикамид	trop	[no SUBSTEM]	tropicamide	atropine derivatives
uridine	[no SUBSTEM]	trifluridine	уридин	[no SUBSTEM]	трифлуридин	uridina	[no SUBSTEM]	trifluridina	uridine derivatives used as antiviral agents and as antineoplastics
	ur	tegafur		ур	тегафур		ur	tegafur	
	vudine	lamivudine zidovudine		вудин	ламивудин зидовудин		vudina	lamivudina zidovudina	
vaptan	[no SUBSTEM]	tolvaptan	ваптан	[no SUBSTEM]	толваптан	vaptan	[no SUBSTEM]	tolvaptan	vasopressin receptor antagonists
verine	[no SUBSTEM]	alverine dicycloverine mebeverine propiverine	верин	[no SUBSTEM]	алверин дицикловерин мебеверин пропиверин	verina	[no SUBSTEM]	alverina dicicloverina mebeverina propiverina	spasmolytics with a papaverine-like action
vin	[no SUBSTEM]	vinblastine vincristine vindesine vinorelbine	вин	[no SUBSTEM]	винбластин винкристин виндезин винорелбин	vin	[no SUBSTEM]	vinblastina vincristina vindesina vinorelbina	vinca alkaloids
vir	[no SUBSTEM]	efavirenz enfuvirtide letermovir nevirapine	вир	[no SUBSTEM]	эфавиренз энфувиртид летермовир невирапин	vir	[no SUBSTEM]	efavirenz enfuvirtide letermovir nevirapina	antivirals (undefined group)

		ribavirin			рибавирин			ribavirina	
	amivir	oseltamivir zanamivir		амивир	осельтамивир занамивир		amivir	oseltamivir zanamivir	
	asvir	daclatasvir elbasvir ledipasvir velpatasvir		асвир	даклатасвир элбасвир ледипасвир велпатасвир		asvir	daclatasvir elbasvir ledipasvir velpatasvir	
	buvir	dasabuvir sofosbuvir		бувир	дасабувир софосбувир		buvir	dasabuvir sofosbuvir	
	cavir	abacavir entecavir		кавир	абакавир энтекавир		cavir	abacavir entecavir	
	ciclovir	aciclovir famciclovir ganciclovir penciclovir valaciclovir valganciclovir		цикловир	ацикловир фамцикловир ганцикловир пенцикловир валацикловир валганцикловир		ciclovir	aciclovir famciclovir ganciclovir penciclovir valaciclovir valganciclovir	
	fovir	adefovir tenofovir		фовир	адефовир тенофовир		fovir	adefovir tenofovir	
	navir	atazanavir darunavir fosamprenavir lopinavir ritonavir saquinavir tipranavir		навир	атазанавир дарунавир фосампренавир лопинавир ритонавир саквинавир типранавир		navir	atazanavir darunavir fosamprenavir lopinavir ritonavir saquinavir tipranavir	

	previr	glecaprevir paritaprevir voxilaprevir		превир	глекапревир паритапревир воксилапревир		previr	glecaprevir paritaprevir voxilaprevir	
	tegravir	bictegravir dolutegravir elvitegravir raltegravir		тегравир	биктегравир долутегравир элвитегравир ралтегравир		tegravir	bictegravir dolutegravir elvitegravir raltegravir	
	virine	doravirine etravirine rilpivirine		вирин	доравирин этравирин рилпивирин		virina	doravirina etravirina rilpivirina	
	viroc	maraviroc		вирок	маравирок		viroc	maraviroc	
xaban	[no SUBSTEM]	apixaban edoxaban rivaroxaban	ксабан	[no SUBSTEM]	апиксабан эдоксабан ривароксабан	xaban	[no SUBSTEM]	apixaban edoxaban rivaroxaban	blood coagulation factor XA inhibitors, antithrombotics
zolid	[no SUBSTEM]	linezolid tedizolid	золид	[no SUBSTEM]	линезолид тедизолид	zolid	[no SUBSTEM]	linezolid tedizolid	oxazolidinone antibacterials
zomib	[no SUBSTEM]	bortezomib carfilzomib ixazomib	зомиб	[no SUBSTEM]	бортезомиб карфилзомиб иксазомиб	zomib	[no SUBSTEM]	bortezomib carfilzomib ixazomib	proteasome inhibitors

A16 Frequent STEM taxa in English and Russian

STEM taxon	Examples of $INNs_{EN}$	Russian STEM	Examples of $INNs_{RU}$	Pharmacological group	#
ac	ketorolac; sulindac	ак	кеторолак; сулиндак	anti-inflammatory agents, ibufenac derivatives	8
arabine/citabine	decitabine fludarabine	арабин; цитабин	децитабин; флударабин	arabinofuranosyl derivatives / nucleosides antiviral or antineoplastic agents, cytarabine or azacitidine derivatives	8
ase	cocarboxylase; idursulfase	аза	кокарбоксилаза; идурсульфаза	enzymes	13
astine	acrivastine; clemastine	астин	акривастин; клемастин	antihistaminics	6
azepam (and SUBSTEMS azolam and azam; variant STEM opam)	clobazam; diazepam; midazolam; nefopam	азепам (and SUBSTEMS азолам and азолам; variant STEM опам)	клобазам; диазепам; мидазолам; нефопам	diazepam derivatives	13
azosin (and variant STEMS ulosin, uzosin)	prazosin; tamsulosin; alfuzosin	азозин (and variant STEMS улозин, узозин)	празозин; тамсулозин; алфузозин		
caine	lidocaine; tetracaine	каин	лидокаин; тетракаин	local anaesthetics	13
calci	calcitriol; ergocalciferol	кальци	кальцитриол; эргокальциферол	vitamin D analogues/derivatives	7
cef	ceftriaxone; cefuroxime	цеф	цефтриаксон; цефуроксим	antibiotics, cefalosporanic acid derivatives	10
cillin	benzylpenicillin; flucloxacillin	циллин	бензилпенициллин; флуклоксациллин	antibiotics, 6-aminopenicillanic acid derivatives	6
conazole	ketoconazole; miconazole	коназол	кетоконазол; миконазол	systemic antifungal agents, miconazole derivatives	10
cycline	doxycycline; oxytetracycline	циклин	доксициклин; окситетрациклин	antibiotics, protein-synthesis inhibitors, tetracycline derivatives	7
dil (and SUBSTEMS dilol and dyl)	bisacodyl; carvedilol trihexyphenidyl	дил	бисакодил; карведилол тригексифенидил	vasodilators	5
dipine	amlodipine; nifedipine	дипин	амлодипин; нифедипин	calcium channel blockers, nifedipine derivatives	6

drine (and SUBSTEMS frine / phrine)	epinephrine; pseudoephedrine	дрин (and SUBSTEM фрин)	эпинефрин; псевдоэфедрин	sympathomimetics	5
gab	gabapentin; tiagabine	габ	габапентин; тиагабин	gabamimetic agents	5
gest	desogestrel; progesterone	гест	дезогестрел; прогестерон	steroids, progestogens	12
gli	pioglitazone; sitagliptin	гли	пиоглитазон; ситаглиптин	antihyperglycaemics	14
grel	anagrelide; ticagrelor	грел	анагрелид; тикагрелор	platelet aggregation inhibitors	5
ifene	clomifene; tamoxifen	ифен	кломифен; тамоксифен	antioestrogens or oestrogen receptor modulators, clomifene and tamoxifen derivatives	6
imus	everolimus; tacrolimus	имус	эверолимус; такролимус	immunosuppressants (other than antineoplastics)	5
izine	cetirizine; hydroxyzine	изин	цетиризин; гидроксизин	diphenylmethyl piperazine derivatives	6
mab	infliximab; trastuzumab	маб	инфликсимаб; трастузумаб	monoclonal antibodies	63
mycin	azithromycin; erythromycin	мицин	азитромицин; эритромицин	antibiotics, produced by Streptomyces strains (see also -kacin)	11
olol (and SUBSTEM alol)	bisoprolol; propranolol; labetalol	олол (and SUBSTEM алол)	бисопролол; пропранолол; лабеталол	β-adrenoreceptor antagonists	16
oxetine	fluoxetine; paroxetine	оксетин	флуоксетин; пароксетин	serotonin and/or norepinephrine reuptake inhibitors, fluoxetine derivatives	7
pine (and SUBSTEMS apine, oxepin; zepine)	clozapine; doxepin	пин	клозапин; доксепин	tricyclic compounds	10
prazole	lansoprazole; omeprazole	празол	лансопразол; омепразол	antiulcer, benzimidazole derivatives	6
pril	captopril; ramipril	прил	каптоприл; рамиприл	angiotensin-converting enzyme inhibitors	10
prost	misoprostol; tafluprost	прост	мизопростол; тафлупрост	prostaglandins	12
rafenib	dabrafenib; sorafenib	рафениб	дабрафениб; сорафениб	Raf (rapidly accelerated fibrosarcoma) kinase inhibitors	5
relin	goserelin; somatorelin	релин	гозерелин; соматорелин	pituitary hormone-release stimulating peptides	7
retin	isotretinoin; retinol	ретин	изотретиноин; ретинол	retinol derivatives	5
sartan	candesartan; valsartan	сартан	кандесартан; валсартан	angiotensin II receptor antagonists, antihypertensive (non-peptidic)	8

stat (and SUBSTEM vastatin)	atorvastatin; telotristat	стат	аторвастатин; телотристат	enzyme inhibitors (statins)	13
stim	filgrastim; pegfilgrastim	стим	филграстим; пэгфилграстим	colony stimulating factors	5
terol	formoterol; salmeterol	терол	формотерол; салметерол	bronchodilators, phenethylamine derivatives	6
tide	lanreotide; semaglutide	тид	ланреотид; семаглутид	peptides and glycopeptides (for special groups of peptides see -actide, -pressin, -relin, -tocin)	16
tinib	erlotinib; upadacitinib	тиниб	эрлотиниб; упадацитиниб	tyrosine kinase inhibitors	29
triptan	sumatriptan; zolmitriptan	триптан	суматриптан; золмитриптан	serotonin (5HT1) receptor agonists, sumatriptan derivatives	7
vir	aciclovir; enfuvirtide; maraviroc	вир	ацикловир; энфувиртид; маравирок	antivirals (undefined group)	41

A17 Top ten beginnings of INNs in Russian and English (initial bigrams and trigrams)

Russian initial ngrams (n = 1046)						English initial ngrams (n = 1046)					
Bigram	**n**	**%**	**Trigram**	**n**	**%**	**Bigram**	**n**	**%**	**Trigram**	**n**	**%**
ме	35	3.3	флу	16	1.5	me-	34	3.3	flu-	18	1.7
ка	28	6.0	про	16	1.5	ca-	31	3.0	pro-	16	1.5
пр	27	8.6	кар	14	1.3	pr-	28	2.7	car-	14	1.3
тр	26	11.1	мет	14	1.3	tr-	24	2.3	met-	14	1.3
те	23	13.3	кло	12	1.1	te-	23	2.2	chl-	12	1.1
ди	22	15.4	три	12	1.1	di-	20	1.9	clo-	12	1.1
да	21	17.4	хло	12	1.1	le-	20	1.9	tri-	11	1.1
ле	21	19.4	лев	11	1.1	be-	19	1.8	cef-	10	1.0
бе	20	21.3	окс	10	1.0	da-	19	1.8	lev-	10	1.0
ми	19	23.1	фен	10	1.0	fl-	19	1.8	hyd-	9	0.9
ци	19	25.0	цеф	10	1.0	mi-	19	1.8			
TOTAL	**261**	**24.9**		**137**	**13.1**	**TOTAL**	**256**	**24.5**		**142**	**13.6**

A18 Russian and English initial bigrams and trigrams, descriptive statistics

Type of ngram		Russian	English
First two letters (initial bigram) (Rel-01)	Range	34	33
	Mean	13.61	12.96
	SD	8.10	8.22
	Unique values	142	154
First three letters (initial trigram) (Rel-02)	Range	15	17
	Mean	3.94	3.75
	SD	3.95	3.96
	Unique values	545	567

A19 Top ten endings of INNs in Russian and English (terminal bigrams and trigrams)

Russian terminal ngrams (n = 1046)						English terminal bigrams (n = 1046)					
Bigram	**n**	**%**	**Trigram**	**n**	**%**	**Bigram**	**n**	**%**	**Trigram**	**n**	**%**
ин	355	33.9	маб	62	5.9	ne	312	29.8	ine	228	21.8
ол	106	10.1	тин	53	5.1	in	128	12.2	ide	83	7.9
ид	87	8.3	лин	40	3.8	de	83	7.9	one	64	6.1
он	72	6.9	цин	40	3.8	ol	77	7.4	mab	63	6.0
аб	62	5.9	зин	38	3.6	ab	63	6.0	nib	38	3.6
иб	52	5.0	ниб	37	3.5	ib	52	5.0	cin	36	3.4
ан	42	4.0	мид	36	3.4	an	35	3.3	vir	33	3.2
ил	37	3.5	дин	35	3.3	il	33	3.2	ole	30	2.9
ир	33	3.2	вир	33	3.2	ir	33	3.2	tin	22	2.1
ам	28	2.7	мин	33	3.2	le	30	2.9	tan	19	1.8
TOTAL	**914**	**87.45**		**397**	**37.94**		**846**	**80.9**		**616**	**58.9**

A20 Russian and English terminal bigrams and trigrams, descriptive statistics

Type of ngram		**Russian**	**English**
Last two letters (terminal bigram) (Rel-03)	Range	354	311
	Mean	156.59	133.63
	SD	144.82	121.35
	Unique values	40	47
Last three letters (terminal trigram) (Rel-04)	Range	61	227
	Mean	23.01	72.21
	SD	18.05	85.81
	Unique values	148	132

A21 Neighbourhood density in Russian and English

	Russian		English	
0 neighbours	100	9.6	134	12.8
Between 1 and 5 neighbours	623	59.6	607	58.0
Between 6 and 10 neighbours	239	22.8	225	21.5
Between 11 and 15 neighbours	50	4.8	44	4.2
Between 16 and 20 neighbours	24	2.3	23	2.2
Between 21 and 25 neighbours	10	1.0	13	1.2

A22 Frequent STEM taxa in English and Italian

STEM taxon	Examples of $INNs_{EN}$	Italian STEM	Examples of Italian names	Pharmacological group	#
ac	ketorolac; sulindac	ac	ketorolac; sulindac	anti-inflammatory agents, ibufenac derivatives	8
arabine/citabine	decitabine; fludarabine	arabina; citabina	fludarabina; decitabina	arabinofuranosyl derivatives / nucleosides antiviral or antineoplastic agents, cytarabine or azacitidine derivatives	8
ase	cocarboxylase; idursulfase	asi	cocarbossilasi; idursulfasi	enzymes	13
astine	acrivastine; clemastine	astina	acrivastina; clemastina	antihistaminics	6
azepam (and SUBSTEMS azolam and azam; variant STEM opam)	clobazam; diazepam; midazolam; nefopam	azepam (and SUBSTEMS azolam and azam; variant STEM opam)	clobazam; diazepam; midazolam; nefopam	diazepam derivatives	13
azosin (and variant STEMS ulosin, uzosin)	prazosin; tamsulosin; alfuzosin	azosina (and variant STEMS ulosina, uzosina)	prazosina; tamsulosina; alfuzosina		
caine	lidocaine; tetracaine	caina	lidocaina; tetracaina	local anaesthetics	13
calci	calcitriol; ergocalciferol	calci	calcitriolo; ergocalciferolo	vitamin D analogues/derivatives	7
cef	ceftriaxone; cefuroxime	cef	ceftriaxone; cefuroxima	antibiotics, cefalosporanic acid derivatives	10
cillin	benzylpenicillin; flucloxacillin	cillina	benzilpenicillina; flucloxacillina	antibiotics, 6-aminopenicillanic acid derivatives	6
conazole	ketoconazole; miconazole	conazolo	ketoconazolo; miconazolo	systemic antifungal agents, miconazole derivatives	10

cycline	doxycycline; oxytetracycline	ciclina	doxiciclina; oxitetraciclina	antibiotics, protein-synthesis inhibitors, tetracycline derivatives	7
dil (and SUBSTEMS dilol and dyl)	bisacodyl; carvedilol; trihexyphenidyl	dil [and SUBSTEMS dilolo and dile]	bisacodil; carvedilolo; triesifenidile	vasodilators	5
dipine	amlodipine; nifedipine	dipina	amlodipina; nifedipina	calcium channel blockers, nifedipine derivatives	6
drine (and SUBSTEM frine, with variant phrine)	epinephrine; pseudoephedrine;	drina (and SUBSTEM frina)	adrenalina; pseudoefedrina	sympathomimetics	5
gab	gabapentin; tiagabine	gab	gabapentin; tiagabina	gabamimetic agents	5
gest	desogestrel; progesterone	gest	desogestrel; progesterone	steroids, progestogens	12
gli	pioglitazone; sitagliptin	gli	pioglitazone; sitagliptin	antihyperglycaemics	14
grel	anagrelide; ticagrelor	grel	anagrelide; ticagrelor	platelet aggregation inhibitors	5
Ifene (and variant STEM ifen)	clomifene; tamoxifen	ifene	clomifene; tamoxifene	antioestrogens or oestrogen receptor modulators, clomifene and tamoxifen derivatives	6
imus	everolimus; tacrolimus	imus	everolimus; tacrolimus	immunosuppressants (other than antineoplastics)	5
izine and variant STEM yzine	cetirizine; hydroxyzine	izina	cetirizina; idroxizina	diphenylmethyl piperazine derivatives	6
mab	infliximab; trastuzumab	mab	infliximab; trastuzumab	monoclonal antibodies	63
mycin	azithromycin; erythromycin	micina	azitromicina; eritromicina	antibiotics, produced by Streptomyces strains (see also -kacin)	11
olol (and SUBSTEM alol)	bisoprolol; propranolol; labetalol	ololo (and SUBSTEM alolo)	bisoprololo; propranololo; labetalolo	β-adrenoreceptor antagonists	16
oxetine	fluoxetine; paroxetine	oxetina	fluoxetina; paroxetina	serotonin and/or norepinephrine reuptake inhibitors, fluoxetine derivatives	7
pine (and SUBSTEMS apine, oxepine, zepine)	clozapine; doxepin	pina (and SUBSTEMS apina, oxepina, zepina)	clozapina; doxepina; eslicarbazepina	tricyclic compounds	10
prazole	lansoprazole; omeprazole	prazolo	lansoprazolo; omeprazolo	antiulcer, benzimidazole derivatives	6
pril	captopril; ramipril	pril	captopril; ramipril	angiotensin-converting enzyme inhibitors	10

prost	misoprostol; tafluprost	prost	misoprostolo; tafluprost	prostaglandins	12
rafenib	dabrafenib; sorafenib	rafenib	dabrafenib; sorafenib	Raf (rapidly accelerated fibrosarcoma) kinase inhibitors	5
relin	goserelin; somatorelin	relina	goserelina; somatorelina	pituitary hormone-release stimulating peptides	7
retin	isotretinoin; retinol	retin	isotretinoina; retinolo	retinol derivatives	5
sartan	candesartan; valsartan	sartan	candesartan; valsartan	angiotensin II receptor antagonists, antihypertensive (non-peptidic)	8
stat (and SUBSTEM vastatin)	atorvastatin; telotristat	stat (and SUBSTEM vastatina)	atorvastatina; telotristat	enzyme inhibitors (statins)	13
stim	filgrastim; pegfilgrastim	stim	filgrastim; pegfilgrastim	colony stimulating factors	5
terol	formoterol; salmeterol	terolo	formoterolo; salmeterolo	bronchodilators, phenethylamine derivatives	6
tide	lanreotide; semaglutide	tide	lanreotide; semaglutide	peptides and glycopeptides (for special groups of peptides see -actide, -pressin, -relin, -tocin)	16
tinib	erlotinib; upadacitinib	tinib	erlotinib; upadacitinib	tyrosine kinase inhibitors	29
triptan	sumatriptan; zolmitriptan	triptan	sumatriptan; zolmitriptan	serotonin (5HT1) receptor agonists, sumatriptan derivatives	7
vir	aciclovir; enfuvirtide; maraviroc	vir	aciclovir; enfuvirtide; maraviroc	antivirals (undefined group)	41

A23 Top ten beginnings of INNs in Italian and English (initial bigrams and trigrams)

Italian initial ngrams (n = 1046)						English initial ngrams (n = 1046)					
Bigram	n	%	Trigram	n	%	Bigram	n	%	Trigram	n	%
me	33	3.2	clo	24	2.3	me-	34	3.3	flu-	18	1.7
cl	28	2.7	flu	18	1.7	ca-	31	3.0	pro-	16	1.5
ca	27	2.6	pro	16	1.5	pr-	28	2.7	car-	14	1.3
pr	27	2.6	met	14	1.3	tr-	24	2.3	met-	14	1.3
tr	26	2.5	car	12	1.1	te-	23	2.2	chl-	12	1.1
te	24	2.3	tri	12	1.1	di-	20	1.9	clo-	12	1.1
fl	20	2.0	cef	10	1.0	le-	20	1.9	tri-	11	1.1
le	20	2.0	fen	10	1.0	be-	19	1.8	cef-	10	1.0
be	19	1.9	lev	10	1.0	da-	19	1.8	lev-	10	1.0
da	19	1.9	ben	8	.8	fl-	19	1.8	hyd-	9	0.9
di	19	1.9	idr	8	.8	mi-	19	1.8			
			tra	8	.8						
TOTAL	**262**	**25.6**		**150**	**14.6**	**TOTAL**	**256**	**24.5**		**142**	**13.6**

A24 Italian and English initial bigrams and trigrams, descriptive statistics

Type of ngram		Italian	English
First two letters (initial bigram) (Rel-01)	Range	32	33
	Mean	13.68	12.96
	SD	8.21	8.22
	Unique values	143	154
First three letters (initial trigram) (Rel-02)	Range	24	17
	Mean	4.12	3.75
	SD	4.87	3.96
	Unique values	547	567

A25 Top ten endings of INNs in Russian and English (terminal bigrams and trigrams)

Italian terminal ngrams (n = 1046)						English terminal bigrams (n = 1046)					
Bigram	**n**	**%**	**Trigram**	**n**	**%**	**Bigram**	**n**	**%**	**Trigram**	**n**	**%**
na	318	30.4	ina	317	30.3	ne	312	29.8	ine	228	21.8
ne	98	9.4	olo	96	9.2	in	128	12.2	ide	83	7.9
lo	96	9.2	ide	82	7.8	de	83	7.9	one	64	6.1
de	82	7.8	one	66	6.3	ol	77	7.4	mab	63	6.0
ab	63	6.0	mab	63	6.0	ab	63	6.0	nib	38	3.6
ib	53	5.1	nib	39	3.7	ib	52	5.0	cin	36	3.4
ir	33	3.2	vir	33	3.2	an	35	3.3	vir	33	3.2
il	32	3.1	ene	21	2.0	il	33	3.2	ole	30	2.9
an	29	2.8	tan	18	1.7	ir	33	3.2	tin	22	2.1
am	27	2.6	ato	12	1.1	le	30	2.9	tan	19	1.8
			ril	12	1.1						
TOTAL	851	83.01		739	72.17		846	80.9		616	58.9

A26 Italian and English terminal bigrams and trigrams, descriptive statistics

Type of ngram		Italian	English
Last two letters (terminal bigram) (Rel-03)	Range	317	311
	Mean	135.45	133.63
	SD	126.142	121.35
	Unique values	54	47
Last three letters (terminal trigram) (Rel-04)	Range	316	227
	Mean	126.49	72.21
	SD	131.036	85.81
	Unique values	114	132

A27 Neighbourhood density in Italian and English

	Italian		English	
0 neighbours	96	9.38	134	12.8
Between 1 and 5 neighbours	559	54.59	607	58.0
Between 6 and 10 neighbours	250	24.41	225	21.5
Between 11 and 15 neighbours	62	6.05	44	4.2
Between 16 and 20 neighbours	43	4.20	23	2.2
Between 21 and 25 neighbours	13	1.27	13	1.2
26+ neighbours	1	0.10	0	0.0

References

1. Abolhassani, N., Sari, A. A., Rashidian, A., & Rastegarpanah, M. (2017). Assessment of orthographic similarity of drugs names between Iran and overseas using the solar model. *Iranian Journal of Public Health*, *46*(12), 1652.
2. Agency, E. M. (2020). *Medication errors - European Medicines Agency*. https://www.ema.europa.eu/en/human-regulatory/post-authorisation/pharmacovigilance/medication-errors, accessed 23.2.2023.
3. Aitken, M., Valkova, S., & Others. (2013). Avoidable costs in US healthcare: the $200 billion opportunity from using medicines more responsibly. *IMS Institute for Healthcare Informatics*, 27–29.
4. Akici, A., Kalaça, S., Uğurlu, M. Ü., Toklu, H. Z., Iskender, E., & Oktay, Ş. (2004). Patient knowledge about drugs prescribed at primary healthcare facilities. *Pharmacoepidemiology and Drug Safety*, *13*(12), 871–876.
5. Allinson, T. T., Szeinbach, S. L., & Schneider, P. J. (2005). Perceived accuracy of drug orders transmitted orally by telephone. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *62*(1), 78–83.

6. Alteren, J., Hermstad, M., White, J., & Jordan, S. (2018). Conflicting priorities: Observation of medicine administration. *Journal of Clinical Nursing*, *27*(19-20), 3613–3621.
7. Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.
8. Andreassen LM, Kjome RLS, Solvik UO, Houghton J, Desborough JA. The potential for deprescribing in care home residents with Type 2 diabetes. International journal of clinical pharmacy. 2016;38(4):977-84.
9. Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461.
10. Anto, B., Barlow, D., Oborne, C. A., & Whittlesea, C. (2011). Incorrect drug selection at the point of dispensing: a study of potential predisposing factors. *The International Journal of Pharmacy Practice*, *19*(1), 51–60. https://doi.org/10.1111/j.2042-7174.2010.00072.x
11. Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (2013). *Critical realism: Essential readings*. Routledge.
12. Aronson, J. K. (2000). Where name and image meet"—the argument for "adrenaline. *BMJ* , *320*(7233), 506–509.
13. Aronson, J. K. (2004). Medication errors resulting from the confusion of drug names. *Expert Opinion on Drug Safety*, *3*(3), 167–172. https://doi.org/10.1517/eods.3.3.167.31069
14. Aronson, J. K. (2009a). Medication errors: definitions and classification. *British Journal of Clinical Pharmacology*, *67*, 599–604.
15. Aronson, J. K. (2009b). Medication errors: EMERGing solutions. *British Journal of Clinical Pharmacology*, *67*(6), 589.
16. Aronson, J. K. (2009c). Medication errors: what they are, how they happen, and how to avoid them. *QJM: An International Journal of Medicine*, *102*(8), 513–521.
17. Aronson, J. K., & Ferner, R. E. (2005). Clarification of terminology in drug safety. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *28*(10), 851–870.

18. Aschenbrenner, D., Boesen, M., Cohen, M. R., Davis, N. M., Esser, L., Felver, L., & Others. (1996). Understanding and preventing medication errors: an education resource. *Rockville (MD): United States Pharmacopeia*.
19. Ashcroft, D. M., Quinlan, P., & Blenkinsopp, A. (2005). Prospective study of the incidence, nature and causes of dispensing errors in community pharmacies. *Pharmacoepidemiology and Drug Safety*, *14*(5), 327–332.
20. ASNA. (2023). *Парацетамол*. https://www.asna.ru/product/paratsetamol/instruction/?utm_referrer=https%3A%2F%2Fwww.google.com%2F
21. Backman, R., Bayliss, S., Moore, D., & Litchfield, I. (2017). Clinical reminder alert fatigue in healthcare: a systematic literature review protocol using qualitative evidence. *Systematic Reviews*, *6*(1), 1–6.
22. Balocco, R., De Sousa Guimaraes Koch, S., Thorpe, R., Weisser, K., & Malan, S. (2022). New INN nomenclature for monoclonal antibodies [Review of *New INN nomenclature for monoclonal antibodies*]. *The Lancet*, *399*(10319), 24. thelancet.com. https://doi.org/10.1016/S0140-6736(21)02732-X
23. Bar-Dayan, Y., Shotashvily, T., Boaz, M., & Wainstein, J. (2017). Error in drugs consumption among older patients. *American Journal of Therapeutics*, *24*(6), e701–e705.
24. Basco, W. T., Ebeling, M., Hulsey, T. C., & Simpson, K. (2010). Using pharmacy data to screen for look-alike, sound-alike substitution errors in pediatric prescriptions. *Academic Pediatrics*, *10*(4), 233–237.
25. Basco, W. T., Garner, S. S., Ebeling, M., Freeland, K. D., Hulsey, T. C., & Simpson, K. (2016). Evaluating the potential severity of look-alike, sound-alike drug substitution errors in children. *Academic Pediatrics*, *16*(2), 183–191.
26. Bates, D. W., Cullen, D. J., Laird, N., Petersen, L. A., Small, S. D., Servi, D., Laffel, G., Sweitzer, B. J., Shea, B. F., Hallisey, R., & Others. (1995). Incidence of adverse drug events and potential adverse drug events: implications for prevention. *JAMA: The Journal of the American Medical Association*, *274*(1), 29–34.
27. Bates, D. W., Leape, L. L., Cullen, D. J., Laird, N., Petersen, L. A., Teich, J. M., Burdick, E., Hickey, M., Kleefield, S., Shea, B., & Others. (1998). Effect of computerized physician order entry and a team intervention on prevention of

serious medication errors. *JAMA: The Journal of the American Medical Association*, *280*(15), 1311–1316.

28. Bates, D. W., Teich, J. M., Lee, J., Seger, D., Kuperman, G. J., Ma'Luf, N., Boyle, D., & Leape, L. (1999). The impact of computerized physician order entry on medication error prevention. *Journal of the American Medical Informatics Association: JAMIA*, *6*(4), 313–321.
29. Bauer, Laurie. 1979. On the need for pragmatics in the study of nominal compounding. Journal of Pragmatics 3(1). 45–50.
30. Bauer, L. (1998). When is a sequence of two nouns a compound in English? *English Language and Linguistics*, *2*(1), 65–86. https://doi.org/10.1017/s1360674300000691
31. Bauer, L. (2007). *Linguistics Student's Handbook*. Edinburgh University Press.
32. Bell, M. (2011). At the boundary of morphology and syntax. *Morphology and Its Interfaces*, *178*, 137–168.
33. Bell, Melanie J. 2012. The English NN construct: Its prosody and structure. University of Cambridge dissertation.
34. Bell, Melanie J. & Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. Journal of Linguistics 48. 485–520.
35. Bell, M. J., & Schäler, M. (2013). *Semantic transparency: challenges for distributional semantics*. 1–10.
36. Bell, M. J., & Schäfer, M. (2016). Modelling semantic transparency. *Morphology*, *26*(2), 157–199.
37. Berman, A. (2004). Reducing medication errors through naming, labeling, and packaging. *Journal of Medical Systems*, *28*(1), 9–29.
38. Blake, N. (1996). *A history of the English language*. Macmillan International Higher Education.
39. Boockvar, K. S., Ho, W., Pruskowski, J., DiPalo, K. E., Wong, J. J., Patel, J., Nebeker, J. R., Kaushal, R., & Hung, W. (2017). Effect of health information exchange on recognition of medication discrepancies is interrupted when data charges are introduced: results of a cluster-randomized controlled trial. *Journal of the American Medical Informatics Association: JAMIA*, *24*(6), 1095–1101.

40. Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.
41. Bracco, F., & Ivaldi, M. (2023). How metaphors of organizational accidents and their graphical representations can guide (or bias) the understanding and analysis of risks. Journal of intelligence, 11(10), 199.
42. Bramley, D. E. P. (2009). The ability of anaesthetists to identify generic medications from trade names. *Anaesthesia and Intensive Care*, *37*(4), 624–629.
43. Brass, E. P., & Weintraub, M. (2003). Label development and the label comprehension study for over-the-counter drugs. *Clinical Pharmacology & Therapeutics*, *74*(5), 406–412.
44. Brodell, R. T., Helms, S. E., KrishnaRao, I., & Bredle, D. L. (1997). Prescription errors. Legibility and drug name confusion. *Archives of Family Medicine*, *6*(3), 296–298.
45. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, *72*(1), 2–10.
46. Brown, C. L., Mulcaster, H. L., Triffitt, K. L., Sittig, D. F., Ash, J. S., Reygate, K., Husband, A. K., Bates, D. W., & Slight, S. P. (2017). A systematic review of the types and causes of prescribing errors generated from using computerized provider order entry systems in primary and secondary care. *Journal of the American Medical Informatics Association: JAMIA*, *24*(2), 432–440.
47. Bryan, R. (2015). Chapter One: Taxonomy and transparency in international pharmaceutical nomenclature. *Word Formation and Transparency in Medical English*, 13.
48. Bryan, R., Aronson, J. K., Boucher, C., Williams, A., & Jordan, S. (2016). PROTOCOL: A systematic review of the literature on look-alike, sound-alike (LASA) medication errors. *PROSPERO*.
49. Bryan, R., Aronson, J. K., ten Hacken, P., Williams, A., & Jordan, S. (2015). Patient safety in medication nomenclature: orthographic and semantic properties of international nonproprietary names. *PloS One*, *10*(12), e0145431.

50. Bryan, R., Aronson, J. K., Williams, A., & Jordan, S. (2020a). The problem of look-alike, sound-alike name errors: Drivers and solutions. *British Journal of Clinical Pharmacology*, *87*(2), 386–394. https://doi.org/10.1111/bcp.14285
51. Bryan, R., Aronson, J. K., Williams, A. J., & Jordan, S. (2020b). A systematic literature review of LASA error interventions. *British Journal of Clinical Pharmacology*, *87*(2), 336–351. https://doi.org/10.1111/bcp.14644
52. Bulfone, G., Romano, F., Puntel, M., & De Lucia, P. (2012). La sicurezza della somministrazione farmacologica in sala operatoria: studio osservazionale. *Professioni Infermieristiche*, *65*(3).
53. Bundy, D. G., Shore, A. D., Morlock, L. L., & Miller, M. R. (2009). Pediatric vaccination errors: application of the “5 rights” framework to a national error reporting database. *Vaccine*, *27*(29), 3890–3896.
54. Cacchiani, S. (2020). (Pseudo-) Anglicisms and Nominal Compounds in Italian. *The Interaction of Borrowing and Word Formation*. https://iris.unimore.it/handle/11380/1219792
55. Campmans, Z., Van Rhijn, A., Dull, R. M., Santen-Reestman, J., Taxis, K., & Borgsteede, S. D. (2018). Preventing dispensing errors by alerting for drug confusions in the pharmacy information system—A survey of users. *PloS One*, *13*(5).
56. Care Quality Commission. (2023). *CQC Insight 14: Medicines safety in NHS trusts*. Care Quality Commission. https://www.cqc.org.uk/publications/major-reports/cqc-insight-14-medicines-safety-nhs-trusts
57. Carney, S. L., Gazarian, M., Denholm, J. T., Reith, D. M., Penhall, R. K., Jenkins, C. R., Wilhelm, K. A., Komesaroff, P. A., Osborn, M. M., & Day, R. O. (2011). What’s in a name? Brand name confusion and generic medicines. *The Medical Journal of Australia*, *195*(11), 650–651.
58. Castle, L., Franzblau-Isaac, E., & Paulsen, J. (2005). Using Six Sigma to reduce medication errors in a home-delivery pharmacy service. *Joint Commission Journal on Quality and Patient Safety / Joint Commission Resources*, *31*(6), 319–324.
59. Chanakit, T., Napaporn, J., Chiempattanakajohn, T., Sangkhawan, S., & Wichakot, S. (2013). The survey of look alike/sound alike (LASA) drugs available in hospitals in Thailand. *African Journal of Pharmacy and Pharmacology*, *7*(6), 227–239.

60. Chan, A. S., Muller, R. J., & Mancuso, C. E. (2018). Description of a tall-man lettering program for look-alike and sound-alike antineoplastic agents. *The Bulletin of the American Society of Hospital Pharmacists*, *75*(8), 512–513.
61. Chandler, D. (2007). *Semiotics: the basics*. Routledge.
62. Cheng, C. M., Salazar, A., Amato, M. G., Lambert, B. L., Volk, L. A., & Schiff, G. D. (2018). Using drug knowledgebase information to distinguish between look-alike-sound-alike drugs. *Journal of the American Medical Informatics Association: JAMIA*, *25*(7), 872–884.
63. Chisholm, C. D., & Klanduch, F. (2000). Inadvertent Administration of Sufentanil Instead of Fentanyl during Sedation/Analgesia in a Community Hospital Emergency Department A Report of Two Cases. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *7*(11), 1282–1284.
64. Chua, S.-S., Wong, I. C. K., Edmondson, H., Allen, C., Chow, J., Peacham, J., Hill, G., & Grantham, J. (2003). A feasibility study for recording of dispensing errors and “near misses” in four UK primary care pharmacies. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *26*(11), 803–813.
65. Cina, J. L., Gandhi, T. K., Churchill, W., Fanikos, J., McCrea, M., Mitton, P., Rothschild, J. M., Featherstone, E., Keohane, C., Bates, D. W., & Others. (2006). How many hospital pharmacy medication dispensing errors go undetected? *Joint Commission Journal on Quality and Patient Safety / Joint Commission Resources*, *32*(2), 73–80.
66. Ciociano, N., & Bagnasco, L. (2014). Look alike/sound alike drugs: a literature review on causes and solutions. *International Journal of Clinical Pharmacy*, *36*(2), 233–242.
67. Cohen, M. R. (1995). Drug product characteristics that foster drug-use-system errors. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *52*(4), 395–399.
68. Cohen MR, S. J. L. (2014). Cardizem-Cardene Mix-up. *Hospital Pharmacy*, *49*, 117–120.
69. Cohen MR, S. J. L. (2015). Alteplase and Tenecteplase Confusion. *Hospital Pharmacy*, *50*, 849–854.

70. Coletti, M. H., & Bleich, H. L. (2001). Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association: JAMIA*, *8*(4), 317–323.
71. Cook, R. I., & Woods, D. D. (1994). Operating at the sharp end: the complexity of human error. *Human Error in Medicine*, *13*, 225–310.
72. Council of Europe. (2023a). *European Pharmacopoeia Online*. Council of Europe. https://pheur.edqm.eu/subhome/11-0
73. Council of Europe. (2023b). *Individual monographs: virtual training programme (European Directorate for the Quality of Medicines and Healthcare)*. Council of Europe. https://www.edqm.eu/documents/52006/789998/Presentation+-+Individual+monographs+on++%E2%80%9CSubstances+for+pharmaceutical+use%E2%80%9D++and+on+Medicinal+Products+%28containing+chemically++defined+APIs%29.pdf/1f213550-8f03-9a8f-4bd1-d1c21fa54c9a?t=1658475557879
74. Cranshaw, J., Gupta, K. J., & Cook, T. M. (2009). Litigation related to drug errors in anaesthesia: an analysis of claims against the NHS in England 1995--2007. *Anaesthesia*, *64*(12), 1317–1323.
75. Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
76. Cubberley, P. (2002). *Russian: A Linguistic Introduction*. Cambridge University Press. https://play.google.com/store/books/details?id=yOGzAyN3V88C
77. Cutter, J., & Jordan, S. (2013). The systems approach to error reduction: factors influencing inoculation injury reporting in the operating theatre. *Journal of Nursing Management*, *21*(8), 989–1000.
78. Dalton, K., O'Brien, G., O'Mahony, D., & Byrne, S. (2018). Computerised interventions designed to reduce potentially inappropriate prescribing in hospitalised older adults: a systematic review and meta-analysis. *Age and Ageing*, *47*(5), 670–678.
79. Dardano, M. (1978). La formazione delle parole nell'italiano di oggi : primi materiali e proposte. *(No Title)*. https://cir.nii.ac.jp/crid/1130282270351053312
80. Dardano, M. (1994). I linguaggi scientifici. *Storia Della Lingua Italiana*, *2*, 497–551.

81. Davies, E. C., Green, C. F., Mottram, D. R., & Pirmohamed, M. (2010). Interpreting adverse drug reaction (ADR) reports as hospital patient safety incidents. *British Journal of Clinical Pharmacology*, *70*(1), 102–108.
82. Davis, N. M., Cohen, M. R., & Teplitsky, B. (1992). Look-alike and sound-alike drug names: the problem and the solution. *Hospital Pharmacy*, *27*(2), 95–98.
83. DeHenau, C., Becker, M. W., Bello, N. M., Liu, S., & Bix, L. (2016). Tallman lettering as a strategy for differentiation in look-alike, sound-alike drug names: the role of familiarity in differentiating drug doppelgangers. *Applied Ergonomics*, *52*, 77–84.
84. Dekker, S. W. A. (2002). *The field guide to human error investigations. Bedford, UK*. Cranfield University Press/Aldershot, UK: Ashgate.
85. Deléger, L., Namer, F., & Zweigenbaum, P. (2009). Morphosemantic parsing of medical compound words: Transferring a French analyzer to English. *International Journal of Medical Informatics*, *78*, S48–S55.
86. Dembicki, E. L. (1967). Psychiatric drugs & trends. Look-alike and sound-alike drugs. *Journal of Psychiatric Nursing and Mental Health Services*, *5*(6), 604–608.
87. Derrida, J. (1982). *Margins of philosophy*. University of Chicago Press.
88. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.
89. Doyle, L., Brady, A.-M., & Byrne, G. (2009). An overview of mixed methods research. *Journal of Research in Nursing: JRN*, *14*(2), 175–185.
90. Drahl, C. (2012). *Where drug names come from. Behind every generic name lies a specific process*. https://cen.acs.org/articles/90/i3/Drug-Names-Come.html
91. Dsouza, S., & Kulkarni, A. (2017). The hazards of look alike packaging in anaesthesia practice. *Asian Journal of Anesthesiology*, *55*(4), 91–92.
92. Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433.
93. Durham, W., Sykes, C., Piper, S., & Stokes, P. (2015). Conceptual frameworks and terminology in doctoral nursing research. *Nurse Researcher*, *23*(2).

94. Elliott, R., Camacho, E., Campbell, F., Jankovic, D., St James, M. M., Kaltenthaler, E., Wong, R., Sculpher, M., & Faria, R. (2018). Prevalence and economic burden of medication errors in the NHS in England. *Rapid Evidence Synthesis and Economic Analysis of the Prevalence and Burden of Medication Error in the UK*.
95. eMC. (2021). *Electronic Medicines Compendium*. https://www.medicines.org.uk/emc
96. Emmerton, L. M., & Rizk, M. F. S. (2012). Look-alike and sound-alike medicines: risks and “solutions.” *International Journal of Clinical Pharmacy*, *34*(1), 4–8.
97. Emmerton, L. M., Rizk, M. F. S., Bedford, G., & Lalor, D. (2015). Systematic derivation of an Australian standard for Tall Man lettering to distinguish similar drug names. *Journal of Evaluation in Clinical Practice*, *21*(1), 85–90.
98. EU Directive. (1992). *Council Directive of 31 March 1992. Official Journal of the European Communities No L. 1992. p. 31. March:113/8-12 (92/27/EEC)*.
99. Evans, S. M., Berry, J. G., Smith, B. J., Esterman, A., Selim, P., O’Shaughnessy, J., & DeWit, M. (2006). Attitudes and barriers to incident reporting: a collaborative hospital study. *BMJ Quality & Safety*, *15*(1), 39–43.
100. Farbstein, K., & Clough, J. (2001). Improving medication safety across a multihospital system. *The Joint Commission Journal on Quality Improvement*, *27*(3), 123–137.
101. Farmer, E. W., & Taylor, R. M. (1980). Visual search through color displays: Effects of target-background similarity and background uniformity. *Perception & Psychophysics*, *27*(3), 267–272.
102. Fasting, S., & Gisvold, S. E. (2000). Adverse drug errors in anesthesia, and the impact of coloured syringe labels. *Canadian Journal of Anaesthesia = Journal Canadien D’anesthesie*, *47*(11), 1060.
103. Feldman, L. B., Soltano, E. G., Pastizzo, M. J., & Francis, S. E. (2004). What do graded effects of semantic transparency reveal about morphological processing? *Brain and Language*, *90*(1-3), 17–30.
104. Ferner, R. E., & Aronson, J. K. (2006). Clarification of terminology in medication errors. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *29*(11), 1011–1022.

105. Ferner, R. E., & Aronson, J. K. (2016). Nominal ISOMERs (Incorrect Spellings Of Medicines Eluding Researchers)—variants in the spellings of drug names in PubMed: a database review. *BMJ* , *355*. https://doi.org/10.1136/bmj.i4854
106. Filik, R., Price, J., Darker, I., Gerrett, D., Purdy, K., & Gale, A. (2010). The Influence of tall man lettering on drug name confusion. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *33*(8), 677–687.
107. Filik, R., Purdy, K., Gale, A., & Gerrett, D. (2004). Drug name confusion: evaluating the effectiveness of capital ("Tall Man") letters using eye movement data. *Social Science & Medicine*, *59*(12), 2597–2601.
108. Filik, R., Purdy, K., Gale, A., & Gerrett, D. (2006). Labeling of medicines and patient safety: evaluating methods of reducing drug name confusion. *Human Factors*, *48*(1), 39–47. https://doi.org/10.1518/001872006776412199
109. Friedman, A. L., Geoghegan, S. R., Sowers, N. M., Kulkarni, S., & Formica, R. N. (2007). Medication errors in the outpatient setting: classification and root cause analysis. *Archives of Surgery*, *142*(3), 278–283.
110. Frisson, Steven, Elizabeth Niswander-Klement & Alexander Pollatsek. 2008. The role of semantic transparency in the processing of English compound words. British Journal of Psychology 99(1). 87–107.
111. Gabriele, S. (2006). The role of typography in differentiating look-alike/sound-alike drug names. *Healthcare Quarterly* , *9*, 88–95.
112. Galanter, W. L., Bryson, M. L., Falck, S., Rosenfield, R., Laragh, M., Shrestha, N., Schiff, G. D., & Lambert, B. L. (2014). Indication alerts intercept drug name confusion errors during computerized entry of medication orders. *PloS One*, *9*(7).
113. Galvin, R. S. (1998). Part II: What do employers mean by" value"? *Integrated Healthcare Report*, 1–10.
114. Gangakhedkar, G. R., Waghalkar, P. V., Shetty, A. N., & Dalvi, A. M. (2019). Look-alike drugs: Avoiding potential medical errors. *International Journal of Preventive Medicine*, *10*(1), 10–19.
115. Garg, A. X., Adhikari, N. K. J., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., Sam, J., & Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance

and patient outcomes: a systematic review. *JAMA: The Journal of the American Medical Association*, *293*(10), 1223–1238.

116. Georgieva, V. (2013). Synonymy in English and Bulgarian Military Terminology. *Armenian Folia Anglistika*, *9*(1-2 (11)), 17–30. https://doi.org/10.46991/afa/2013.9.1-2.017
117. Giegerich, H. J. (2004). Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics*, *8*(1), 1–24. https://doi.org/10.1017/s1360674304001224
118. Goldberg, A. (1987). Introduction. In R. D. Mann (Ed.), *Adverse drug reactions: The scale and nature of the problem and the way forward* (pp. 1–3). Parthenon Publishing.
119. Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*(3), 518.
120. Griffith, R., Griffiths, H., & Jordan, S. (2003). Administration of medicines part 1: the law and nursing. *Nursing Standard: Official Newspaper of the Royal College of Nursing*, *18*(2), 47–56.
121. Guevara, E., & Scalise, S. (2009). Searching for Universals in Compounding. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (pp. 101–128). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8825-4_6
122. Harlow, L. L. (2014). *The essence of multivariate thinking: Basic themes and methods*. Routledge.
123. Harré, R. (1987). Enlarging the paradigm. *New Ideas in Psychology*, *5*(1), 3–12.
124. Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, *35*(4), 579-592.
125. Her, Q. L., & Toh, S. (2019). Sound-Alike Look-Alike Confusion and Matching Medication Product Attributes: Simulated Case-Control Studies. *The Annals of Pharmacotherapy*, *53*(10), 973–980.
126. Hoffman, J. M., & Proulx, S. M. (2003). Medication errors caused by confusion of drug names. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *26*(7), 445–452.

127. Iacobini, C. (2004). Composizione con elementi neoclassici. *La Formazione Delle Parole in Italiano. Tübingen: Max Niemeyer Verlag*, 69–95.
128. Irwin, A., Mearns, K., Watson, M., & Urquhart, J. (2011). Pharmacist dispensing error: The effect of neighbourhood density and time pressure on accurate visual perception of drug names. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*, 1621–1625.
129. Irwin, A., Mearns, K., Watson, M., & Urquhart, J. (2013). The effect of proximity, Tall Man lettering, and time pressure on accurate visual perception of drug names. *Human Factors*, *55*(2), 253–266.
130. Irwin, A., Ross, J., Seaton, J., & Mearns, K. (2011). Retrospective analysis of DATIX dispensing error reports from Scottish NHS hospitals. *The International Journal of Pharmacy Practice*, *19*(6), 417–423.
131. ISMP. (2019). *ISMP confused drug name list*. https://www.ismp.org/recommendations/confused-drug-names-list
132. Jackendoff, R. (1983). *Semantics and cognition* (Vol. 8). MIT press.
133. James, K. L., Barlow, D., McArtney, R., Hiom, S., Roberts, D., & Whittlesea, C. (2009). Incidence, type and causes of dispensing errors: a review of the literature. *The International Journal of Pharmacy Practice*, *17*(1), 9–30.
134. Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265–276.
135. Jennings, J., & Foster, J. (2007). Medication safety: just a label away. *AORN Journal*, *86*(4), 618–625.
136. Jordan, S., & Kyriacos, U. (2014). Medicines' management: a public health problem on nursing's agenda. *Journal of Nursing Management*, *22*(3), 271–275.
137. Joseph, J. E. (2004). *The linguistic sign* (Vol. 75). Cambridge: Cambridge University Press.
138. Joshi, M. C., Joshi, H. S., Tariq, K., Ejaj, A., Prayag, S., & Raju, A. (2007). A prospective study of medication errors arising out of look-alike and sound-alike brand names confusion. *The International Journal of Risk & Safety in Medicine*, *19*(4), 195–201.
139. Jurafsky, D., & Martin, J. (2000). *Speech and Language processing*. New Jersey: Prentice-Hall, Inc.

140. Kaushal, R., Bates, D. W., Landrigan, C., McKenna, K. J., Clapp, M. D., Federico, F., & Goldmann, D. A. (2001). Medication errors and adverse drug events in pediatric inpatients. *JAMA: The Journal of the American Medical Association*, *285*(16), 2114–2120.
141. Kaushal, R., Jha, A. K., Franz, C., Glaser, J., Shetty, K. D., Jaggi, T., Middleton, B., Kuperman, G. J., Khorasani, R., Tanasijevic, M., & Others. (2006). Return on investment for a computerized physician order entry system. *Journal of the American Medical Informatics Association: JAMIA*, *13*(3), 261–266.
142. Kenagy, J. W., & Stein, G. C. (2001). Naming, labeling, and packaging of pharmaceuticals. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *58*(21), 2033–2041.
143. Kondrak, G., & Dorr, B. (2006). Automatic identification of confusable drug names. *Artificial Intelligence in Medicine*, *36*(1), 29–42.
144. Koppel, R., Wetterneck, T., Telles, J. L., & Karsh, B.-T. (2008). Workarounds to barcode medication administration systems: their occurrences, causes, and threats to patient safety. *Journal of the American Medical Informatics Association: JAMIA*, *15*(4), 408–423.
145. Kovacic, L., & Chambers, C. (2011). Look-alike, sound-alike drugs in oncology. *Journal of Oncology Pharmacy Practice: Official Publication of the International Society of Oncology Pharmacy Practitioners*, *17*(2), 104–118.
146. Lambert, B. L. (1997). Predicting look-alike and sound-alike medication errors. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *54*(10), 1161–1171.
147. Lambert, B. L., Bhaumik, R., Zhao, W., & Bhaumik, D. K. (2016). Detection and prediction limits for identifying highly confusable drug names from experimental data. *Journal of Biopharmaceutical Statistics*, *26*(2), 365–385.
148. Lambert, B. L., Chang, K.-Y., & Gupta, P. (2003). Effects of frequency and similarity neighborhoods on pharmacists' visual perception of drug names. *Social Science & Medicine*, *57*(10), 1939–1955.
149. Lambert, B. L., Chang, K.-Y., & Lin, S.-J. (2001a). Descriptive analysis of the drug name lexicon. *Drug Information Journal*, *35*(1), 163–172.

150. Lambert, B. L., Chang, K.-Y., & Lin, S.-J. (2001b). Effect of orthographic and phonological similarity on false recognition of drug names. *Social Science & Medicine*, *52*(12), 1843–1857.
151. Lambert, B. L., Chang, K.-Y., & Lin, S.-J. (2003). Immediate free recall of drug names: effects of similarity and availability. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *60*(2), 156–168.
152. Lambert, B. L., Dickey, L. W., Fisher, W. M., Gibbons, R. D., Lin, S.-J., Luce, P. A., McLennan, C. T., Senders, J. W., & Clement, T. Y. (2010). Listen carefully: the risk of error in spoken medication orders. *Social Science & Medicine*, *70*(10), 1599–1608.
153. Lambert, B. L., Donderi, D., & Senders, J. W. (2002). Similarity of drug names: comparison of objective and subjective measures. *Psychology & Marketing*, *19*(7-8), 641–661.
154. Lambert, B. L., Galanter, W., Liu, K. L., Falck, S., Schiff, G., Rash-Foanio, C., Schmidt, K., Shrestha, N., Vaida, A. J., & Gaunt, M. J. (2019). Automated detection of wrong-drug prescribing errors. *BMJ Quality & Safety*, *28*(11), 908–915.
155. Lambert, B. L., Lin, S.-J., Chang, K.-Y., & Gandhi, S. K. (1999). Similarity as a risk factor in drug-name confusion errors: the look-alike (orthographic) and sound-alike (phonetic) model. *Medical Care*, *37*(12), 1214–1225.
156. Lambert, B. L., Lin, S.-J., & Tan, H. (2005). Designing safe drug names. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *28*(6), 495–512.
157. Lambert, B. L., Schroeder, S. R., Cohen, M. R., & Paparella, S. (2022). Beyond mixed case lettering: reducing the risk of wrong drug errors requires a multimodal response. *BMJ Quality & Safety*. https://doi.org/10.1136/bmjqs-2022-014841
158. Lambert, B. L., Schroeder, S. R., & Galanter, W. L. (2016). *Does Tall Man lettering prevent drug name confusion errors? Incomplete and conflicting evidence suggest need for definitive study*.
159. Lambert, B. L., Yu, C., & Thirumalai, M. (2004). A system for multiattribute drug product comparison. *Journal of Medical Systems*, *28*(1), 31–56.

160. Lang, T., & Altman, D. (2013). Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In P. Smart, H. Maisonneuve, & A. (eds) Polderman (Eds.), *Science Editors' Handbook*. European Association of Science Editors.
161. Larmené-Beld, K. H. M., Alting, E. K., & Taxis, K. (2018). A systematic literature review on strategies to avoid look-alike errors of labels. *European Journal of Clinical Pharmacology*, *74*(8), 985–993.
162. Lesar, T. S., Briceland, L., & Stein, D. S. (1997). Factors related to errors in medication prescribing. *JAMA: The Journal of the American Medical Association*, *277*(4), 312–317.
163. Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, *61*(1), 30–44.
164. Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, *84*(1), 50–64.
165. Lisby, M., Nielsen, L. P., Brock, B., & Mainz, J. (2010). How are medication errors defined? A systematic literature review of definitions and characteristics. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care / ISQua*, *22*(6), 507–518.
166. Liu, K., Or, C. K. L., & Li, S. Y. W. (2019). Visual differentiation and recognition memory of look-alike drug names: effects of disfluent format, text enhancement and exposure time. *Ergonomics*, *62*(10), 1289–1300.
167. Li, Y., & Thimbleby, H. (2014). Hot cheese: a processed Swiss cheese model. *JR Coll Physicians Edinb*, *44*, 116–121.
168. Licht, M. H. (1995). Multiple regression and correlation.
169. Lizano-Díez, I., Figueiredo-Escribá, C., Piñero-López, M. Á., Lastra, C. F., Mariño, E. L., & Modamio, P. (2020). Prevention strategies to identify LASA errors: building and sustaining a culture of patient safety. *BMC Health Services Research*, *20*(1), 1–5.
170. Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1.
171. Lüdeling, A. (2005). *Neoclassical word-formation*.

172. Magris, M., & Others. (1992). *La traduzione del linguaggio medico: analisi contrastiva di testi in lingua italiana, inglese e tedesca*.
173. Marecková, E., & Cerven\`y, L. (2002). Latin as the language of medical terminology: some remarks on its role and prospects. *Swiss Medical Weekly*, *132*(4142).
174. Marelli, M., & Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory and Language*, *66*(4), 644–664. https://doi.org/10.1016/j.jml.2012.01.003
175. Marx, D. (1997). Discipline: the role of rule violations. *Ground Effects*, *2*(4), 1–4.
176. McBride, W. G. (1961). Thalidomide and Congenital Abnormalities. *Lancet*, *58*(7), 493–497. https://www.ncbi.nlm.nih.gov/pubmed/19994414
177. McConkie, G. W., & Zola, D. (1979). Is visual information integrated across successive fixations in reading? *Perception & Psychophysics*, *25*(3), 221–224.
178. Mcdougall, D. J., Hoehns, J. D., Feller, T. T., Kriener, S. J., & Witry, M. J. (2016). Inclusion of salt form on prescription medication labeling as a source of patient confusion: a pilot study. *Pharmacy Practice*, *14*(1), 0–0.
179. McNulty, H., & Spurr, P. (1979). Drug names that look or sound alike. *British Medical Journal*, *2*(6194), 836. https://doi.org/10.1136/bmj.2.6194.836
180. MHPRA. (2010). *Mercaptamine and mercaptopurine: confusion between drug names*. https://www.gov.uk/drug-safety-update/mercaptamine-and-mercaptopurine-confusion-between-drug-names
181. Mokdad, M., Mebarki, B., Mebarki, I. E., & Mokdad, I. (2021, May). Dealing with Medical Errors: A Human Factors/Ergonomics Approach. In Congress of the International Ergonomics Association (pp. 389-396). Cham: Springer International Publishing.
182. Moorman, S. M., & Carr, D. (2008). Spouses' effectiveness as end-of-life health care surrogates: accuracy, uncertainty, and errors of overtreatment or undertreatment. *The Gerontologist*, *48*(6), 811–819.
183. Namer, F., & Zweigenbaum, P. (2004). Acquiring meaning for French medical terminology: contribution of morphosemantics. *Medinfo*, 535–539.

184. Neo, H., Ng, S., & Wang, N. (2014). Reduction of medication picking errors in Clementi polyclinic pharmacy by 50% in 6 months. *Annals of the Academy of Medicine, Singapore*, *1*(S265).
185. Newbigging, P. L. (1961). The perceptual redintegration of frequent and infrequent words. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *15*(3), 123.
186. Newcomer, L. N. (2000). Perspective: Medicare Pharmacy Coverage: Ensuring Safety Before Funding: Prescription-related deaths far outnumber other accidental deaths in areas subject to federal oversight (such as cars and airplanes). Why not intervention to ensure patients' safety? *Health Affairs*, *19*(2), 59–62.
187. NHS. (2018a). *NaPSIR data summary: March 2020 (individual CSV files), specifically the "Harm by Type" data table*.
188. NHS. (2018b). *NAPSIR: September 2018 report of incidents reported as having taken place between April 2017 and March 2018*. https://improvement.nhs.uk/documents/3266/NAPSIR_commentary_FINAL_data_to_March_2018.pdf
189. NHS. (2023). *The National Patient Safety Improvement Programmes*. NHS. https://www.england.nhs.uk/patient-safety/patient-safety-improvement-programmes/
190. NHS Digital. (2019a). *Prescription Cost Analysis England, 2019*. https://www.nhsbsa.nhs.uk/statistical-collections/prescription-cost-analysis-england/prescription-cost-analysis-england-2019
191. NHS Digital. (2019b). *Secondary Care Medicines Data (SCMD)*. https://opendata.nhsbsa.net/dataset/secondary-care-medicines-data
192. NICE. (2024, March 25). *Discussing and planning medicines support*. National Institute for Clinical Excellence. https://www.nice.org.uk/about/nice-communities/social-care/quick-guides/discussing-and-planning-medicines-support
193. Nida, E. A. (1977). The Nature of Dynamic Equivalence in Translating. *Babel: International Journal of Translation*. https://eric.ed.gov/?id=EJ177827
194. Nott, M. R. (2001). Misidentification, in-filling and confirmation bias. *Anaesthesia*, *56*(9), 906–924.

195. Nuckols, T. K., Smith-Spangler, C., Morton, S. C., Asch, S. M., Patel, V. M., Anderson, L. J., Deichsel, E. L., & Shekelle, P. G. (2014). The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. *Systematic Reviews*, *3*(1), 56.
196. Nute, C. (2014). Reducing medication errors. *Nursing Standard* , *29*(12), 45.
197. O'Gorman, K. D., & MacIntosh, R. (2014). *Research methods for business and management*. Goodfellow Publishers Limited.
198. Onatade R, Sawieres S, Veck A, Smith L, Gore S, Al-Azeib S. The incidence and severity of errors in pharmacist-written discharge medication orders. International journal of clinical pharmacy. 2017;39(4):722-8.
199. ONS (2023). *Population estimates for England and Wales 2023.* Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/populationestimatesforenglandandwales/mid2023
200. ONS. (2024, March 24). *Taxonomy Oversight Group*. Office for National Statistics. https://www.ons.gov.uk/methodology/classificationsandstandards/taxonomyoversightgroup
201. Onweugbuzie, A. J. (2002). WHY CAN'T WE ALL GET ALONG? TOWARDS A FRAMEWORK FOR UNIFYING RESEARCH PARADIGMS. *Education*, *122*(3).
202. Or, C. K. L., & Chan, A. H. S. (2014). Effects of text enhancements on the differentiation performance of orthographically similar drug names. *Work* , *48*(4), 521–528.
203. Or, C. K. L., & Wang, H. (2014). A comparison of the effects of different typographical methods on the recognizability of printed drug names. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, *37*(5), 351–359.
204. Ostini, R., Roughead, E. E., Kirkpatrick, C. M. J., Monteith, G. R., & Tett, S. E. (2012). Quality Use of Medicines--medication safety issues in naming; look-alike, sound-alike medicine names. *The International Journal of Pharmacy Practice*, *20*(6), 349–357.

205. Oxford English Dictionary [OED]. (2022). *OED*. https://www.oed.com/
206. Pairman, L., Chin, P., Gardiner, S. J., & Doogue, M. (2024). Compulsory indications in hospital prescribing software tested with antibacterial prescriptions. AMIA Summits on Translational Science Proceedings, 2024, 632.
207. Pathak, A., Senard, J. M., Bujaud, T., Bagheri, H., Lapeyre-Mestre, M., Tressieres, M. C., & Montastruc, J. L. (2004). Medication error caused by confusing drug blisters. *The Lancet*, *363*(9427), 2142.
208. Pazour, J. A., & Meller, R. D. (2012). A multiple-drawer medication layout problem in automated dispensing cabinets. *Health Care Management Science*, *15*(4), 339–354.
209. Penn State (2025). *STAT 462 – Lesson 10.7 - Detecting Multicollinearity Using Variance Inflation Factors.* https://online.stat.psu.edu/stat462/node/180/, accessed 3.2.25.
210. Perea, M., & Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica*, *21*(2), 327–340.
211. Perea, M., Rosa, E., & Marcet, A. (2017). Where is the locus of the lowercase advantage during sentence reading? *Acta Psychologica*, *177*, 30–35.
212. PharmaStarTV. (2015, January 11). *Paracetamolo, meccanismo d'azione e applicazioni nel trattamento del dolore*. Youtube. https://www.youtube.com/watch?v=POjFCT9rZro
213. Plag, I. (2018). *Word-formation in English*. Cambridge University Press.
214. Propp, D. A. (2000). Sound-alike and Look-alike Drugs An Overdue Opportunity to Reduce Medical Errors. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *7*(11), 1334–1335.
215. Puato, D. (2011). *Lessico medico e traduzione. Considerazioni contrastive per il tedesco e l'italiano*. https://www.openstarts.units.it/handle/10077/9182
216. Puato, D. (2013). La formazione delle parole nei linguaggi di specializzazione in tedesco: una prospettiva didattica per il lessico medico. *Lingue E Linguaggi*, *8*, 131–144. https://doi.org/10.1285/i22390359v8p131

217. Raban, M. Z., & Westbrook, J. I. (2014). Are interventions to reduce interruptions and errors during medication administration effective?: a systematic review. *BMJ Quality & Safety*, *23*(5), 414–421.
218. Rash-Foanio, C., Galanter, W., Bryson, M., Falck, S., Liu, K. L., Schiff, G. D., Vaida, A., & Lambert, B. L. (2017). Automated detection of look-alike/sound-alike medication errors. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *74*(7), 521–527. https://doi.org/10.2146/ajhp150690
219. Ratwani, R. M., Fairbanks, R. J., Hettinger, A. Z., & Benda, N. C. (2015). Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. *Journal of the American Medical Informatics Association: JAMIA*, *22*(6), 1179–1182.
220. Reason, J. (1990). *Human error*. Cambridge university press.
221. Reason, J. (2000). Human error: models and management. *BMJ* , *320*(7237), 768–770.
222. Reason, J., Hollnagel, E., & Paries, J. (2006). Revisiting the Swiss cheese model of accidents. *Journal of Clinical Engineering*, *27*(4), 110–115.
223. Reinhardt, H., Otte, P., Eggleton, A. G., Ruch, M., Wöhrl, S., Ajayi, S., Duyster, J., Jung, M., Hug, M. J., & Engelhardt, M. (2019). Avoiding chemotherapy prescribing errors: analysis and innovative strategies. *Cancer*, *125*(9), 1547–1557.
224. Ritchie, L. (2013). *Metaphor*. Cambridge University Press. https://play.google.com/store/books/details?id=UoghAwAAQBAJ
225. Roethlisberger, F. J., & Dickson, W. J. (2003). *Management and the Worker* (Vol. 5). Psychology press.
226. Rogers, A. E., Hwang, W.-T., Scott, L. D., Aiken, L. H., & Dinges, D. F. (2004). The working hours of hospital staff nurses and patient safety. *Health Affairs*, *23*(4), 202–212.
227. Rosenthal, R., & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, *19*(1), 115–118.
228. Ross, D. (2004). *Profili morfologici della lingua medica: contrasti in ambito germanico-romanzo*.
229. Rouayroux, N., Calmels, V., Bachelet, B., Sallerin, B., & Divol, E. (2019). Medication prescribing errors: a pre-and post-computerized physician order

entry retrospective study. *International Journal of Clinical Pharmacy*, *41*(1), 228–236.

230. Roumeliotis, N., Sniderman, J., Adams-Webber, T., Addo, N., Anand, V., Rochon, P., Taddio, A., & Parshuram, C. (2019). Effect of Electronic Prescribing Strategies on Medication Error and Harm in Hospital: a Systematic Review and Meta-analysis. *Journal of General Internal Medicine*, 1–14.
231. Routledge, P. (1998). 150 years of pharmacovigilance. *The Lancet*, *351*(9110), 1200–1201. https://doi.org/10.1016/S0140-6736(98)03148-1
232. Runciman, W. B., Roughead, E. E., Semple, S. J., & Adams, R. J. (2003). Adverse drug events and medication errors in Australia. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care / ISQua*, *15*(suppl_1), i49–i59.
233. Rupprecht, D. P., & Rhodes, M. E. (2014). Affixes are essential tools when teaching and studying pharmacological drug classes. *Pharmacy Education*, *14*, 109–115.
234. Ruutiainen, H. K., Kallio, M. M., & Kuitunen, S. K. (2021). Identification and safe storage of look-alike, sound-alike medicines in automated dispensing cabinets. European Journal of Hospital Pharmacy, 28(e1), e151-e156.
235. Sager, J. C. (1990). *Practical course in terminology processing*. John Benjamins Publishing.
236. Salute, L. V. (2020, October 25). *Quali sono gli effetti collaterali comuni dell'omeprazolo?* Youtube. https://www.youtube.com/watch?v=NwXAsukjT-k
237. Sauberan, J. B., Dean, L. M., Fiedelak, J., & Abraham, J. A. (2010). Origins of and solutions for neonatal medication-dispensing errors. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *67*(1), 49–57.
238. Schäfer, M. (2018). *The semantic transparency of English compound nouns*. Language Science Press.
239. Schell, K. L. (2009). Using enhanced text to facilitate recognition of drug names: evidence from two experimental studies. *Applied Ergonomics*, *40*(1), 82–90.
240. Schnoor, J., Rogalski, C., Frontini, R., Engelmann, N., & Heyde, C.-E. (2015). *Case report of a medication error by look-alike packaging*.

241. Schroeder, S. R., Salomon, M. M., Galanter, W. L., Schiff, G. D., Vaida, A. J., Gaunt, M. J., Bryson, M. L., Rash, C., Falck, S., & Lambert, B. L. (2017). Cognitive tests predict real-world errors: the relationship between drug name confusion rates in laboratory-based memory and perception tests and corresponding error rates in large pharmacy chains. *BMJ Quality & Safety*, *26*(5), 395–407.
242. Schulmeister, L. (2006). Look-alike, sound-alike oncology medications. *Clinical Journal of Oncology Nursing*, *10*(1), 35.
243. Segura-Bedmar, I., Martínez, P., & Segura-Bedmar, M. (2008). Drug name recognition and classification in biomedical texts. A case study outlining approaches underpinning automated systems. *Drug Discovery Today*, *13*(17-18), 816–823. https://doi.org/10.1016/j.drudis.2008.06.001
244. Selkirk, Elisabeth O. 1982. The syntax of words. Cambridge, Massachusetts: MIT Press
245. Seoane-Vazquez, E., Rodriguez-Monguio, R., Alqahtani, S., & Schiff, G. (2017). Exploring the potential for using drug indications to prevent look-alike and sound-alike drug errors. *Expert Opinion on Drug Safety*, *16*(10), 1103–1109. https://doi.org/10.1080/14740338.2017.1358361
246. Shah, M. B., Merchant, L., Chan, I. Z., & Taylor, K. (2017). Characteristics that may help in the identification of potentially confusing proprietary drug names. *Therapeutic Innovation & Regulatory Science*, *51*(2), 232–236.
247. Shapiro, L. G., Chung, E., Detwiler, L. T., Mejino, J. L. V., Jr, Agoncillo, A. V., Brinkley, J. F., & Rosse, C. (2005). Processes and problems in the formative evaluation of an interface to the Foundational Model of Anatomy knowledge base. *Journal of the American Medical Informatics Association: JAMIA*, *12*(1), 35–46. https://doi.org/10.1197/jamia.M1401
248. Shapiro, M. (1986). The Russian system of stress. *Russian Linguistics*, *10*(2), 183–204. https://doi.org/10.1007/bf02551662
249. Shappell, S. A., & Wiegmann, D. A. (2000). *The human factors analysis and classification system--HFACS*.
250. Shawahna, R., Rahman, N.-U., Ahmad, M., Debray, M., Yliperttula, M., & Declèves, X. (2013). Impact of prescriber's handwriting style and nurse's duty duration on the prevalence of transcription errors in public hospitals. *Journal of Clinical Nursing*, *22*(3-4), 550–558.

251. Shaw, K. N., Lillis, K. A., Ruddy, R. M., Mahajan, P. V., Lichenstein, R., Olsen, C. S., Chamberlain, J. M., Network, P. E. C. A. R., & Others. (2013). Reported medication events in a paediatric emergency research network: sharing to improve patient safety. *Emergency Medicine Journal: EMJ*, *30*(10), 815–819.
252. Sheskin, D. J. (2007). Handbook Of Parametric And Nonparametric Statistical Procedures.
253. Shiima, Y., Malik, M., & Okorie, M. (2022). Medication Without Harm: Developing optimal medication error reporting systems. Current Drug Safety, 17(1), 7-12.
254. Siebert, J. N., Ehrler, F., Combescure, C., Lovis, C., Haddad, K., Hugon, F., Luterbacher, F., Lacroix, L., Gervaix, A., Manzano, S., & Others. (2019). A mobile device application to reduce medication errors and time to drug delivery during simulated paediatric cardiopulmonary resuscitation: a multicentre, randomised, controlled, crossover trial. *The Lancet Child & Adolescent Health*, *3*(5), 303–311.
255. Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., & Tang, P. C. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association: JAMIA*, *8*(6), 527–534.
256. Slattery, T. J., Schotter, E. R., Berry, R. W., & Rayner, K. (2011). Parafoveal and foveal processing of abbreviations during eye fixations in reading: making a case for case. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(4), 1022.
257. Smither, J. A.-A., & Braun, C. C. (1994). Readability of prescription drug labels by older and younger adults. *Journal of Clinical Psychology in Medical Settings*, *1*(2), 149–159.
258. Smith, H. S., & Lesar, T. S. (2011). Analgesic prescribing errors and associated medication characteristics. *The Journal of Pain: Official Journal of the American Pain Society*, *12*(1), 29–40.
259. SoINN. (2020a). *Course: Linguistic aspects of INN*. https://extranet.who.int/soinn/course/view.php?id=14
260. SoINN. (2020b). *Course: WHO INN Programme - General overview -*. https://extranet.who.int/soinn/course/view.php?id=5

261. Statham, N. (2003). Dynamic Equivalence and Functional Equivalence: How Do They Differ? *The Bible Translator*, *54*(1), 102–111. https://doi.org/10.1177/026009350305400102
262. Stegemann, S., Riedl, R., & Sourij, H. (2017). Identification of different shapes, colors and sizes of standard oral dosage forms in diabetes type 2 patients—A pilot study. *International Journal of Pharmaceutics*, *517*(1-2), 112–118.
263. Stevens, R., Goble, C. A., & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, *1*(4), 398–414.
264. Ten Hacken, P. (1994). *Defining Morphology: A principled approach to determining the boundaries of compounding, derivation, and inflection* [Universität Basel].
265. Ten Hacken, P. (2010). Creating legal terms: A linguistic perspective. *International Journal for the Semiotics of Law-Revue Internationale de Sémiotique Juridique*, *23*(4), 407–425.
266. Ten Hacken, P. (2015). Terms and specialised vocabulary. *Handbook of Terminology.*, 3–13. https://books.google.com/books?hl=en&lr=&id=MQZoBwAAQBAJ&oi=fnd&pg=PA3&dq=terms+and+specialised+vocabularies+handbook+of+terminlogy+ten+hacken&ots=KpiBYK9NcV&sig=AZ1N0D8azcrA7TYpZUA8eGS0aAU
267. Ten Hacken, P., & Panocová, R. (2011). Individual and social aspects of word formation. *Kwartalnik Neofilologiczny*, *3*.
268. Ten Hacken, P., & Panocová, R. (2015). *Word Formation and Transparency in Medical English*. Cambridge Scholars Publishing.
269. Teplitsky, B. (1973). Hazards of sound-alike, look-alike drug names. *California Medicine*, *119*(5), 62.
270. Teplitsky, B. (1985). Look alike/sound alike drug names. *Colorado Nurse* , *95*(4), 27–27.
271. Thimbleby, H., Blandford, A., Cairns, P., Curzon, P., & Jones, M. (2002). User interface design as systems design. In *People and Computers XVI-Memorable Yet Invisible* (pp. 281–301). Springer.
272. Thompson, C. A. (2008). *USP says thousands of drug names look or sound alike*. Oxford University Press.

273. Tseng, H.-Y., Wen, C.-F., Lee, Y.-L., Jeng, K.-C., & Chen, P.-L. (2018). Dispensing errors from look-alike drug trade names. *European Journal of Hospital Pharmacy. Science and Practice*, *25*(2), 96–99.
274. Tsuji, T., Irisa, T., Tagawa, S., Kawashiri, T., Ikesue, H., Kokubu, C., Kanaya, A., Egashira, N., & Masuda, S. (2015). Differences in recognition of similar medication names between pharmacists and nurses: a retrospective study. *Journal of Pharmaceutical Health Care and Sciences*, *1*(1), 19.
275. Tuohy, N., & Paparella, S. (2005). Look-alike and sound-alike drugs: errors just waiting to happen. *Journal of Emergency Nursing: JEN: Official Publication of the Emergency Department Nurses Association*, *31*(6), 569–571.
276. Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.
277. US Food and Drug Administration. (2001). *FDA Medication Error Reporting Program [MERP]*. https://www.pbm.va.gov/vacenterformedicationsafety/othervasafetyprojects/appendixiiusplasa.pdf
278. US Food and Drug Administration. (2007). *Update on Phonetic and Orthographic Computer Analysis Tool*. https://www.fda.gov/drugs/medication-errors-related-cder-regulated-drug-products/update-phonetic-and-orthographic-computer-analysis-tool
279. US Food and Drug Administration. (2020). *Name Differentiation Project*. https://www.fda.gov/drugs/medication-errors-related-cder-regulated-drug-products/fda-name-differentiation-project
280. van den Bemt, P. M., & Egberts, A. C. G. (2007). Drug related problems: definitions and classification. *EJHP Practice*, *2007*(13), 62–64.
281. Van de Vreede, M., McRae, A., Wiseman, M., & Dooley, M. J. (2008). Successful introduction of Tallman letters to reduce medication selection errors in a hospital network. *Journal of Pharmacy Practice and Research*, *38*(4), 263–266.
282. Velo, G. P., & Minuz, P. (2009). Medication errors: prescribing faults and prescription errors. *British Journal of Clinical Pharmacology*, *67*(6), 624–628. https://doi.org/10.1111/j.1365-2125.2009.03425.x
283. Venuti, L. (2017). *The Translator's Invisibility: A History of Translation*. Routledge. https://play.google.com/store/books/details?id=xW5ADwAAQBAJ

284. Wallace, W. P., Stewart, M. T., Shaffer, T. R., & Wilson, J. A. (1998). Are false recognitions influenced by prerecognition processing? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*(2), 299.
285. Wallace, W. P., Stewart, M. T., Sherman, H. L., & Mellor, M. D. (1995). False positives in recognition memory produced by cohort activation. *Cognition*, *55*(1), 85–113.
286. Wang, H., & Or, C. K. L. (2019). Effects of text enhancement, identical prescription-package names, visual cues, and verbal provocation on visual searches of look-alike drug names: a simulation and eye-tracking study. *Human Factors*, 0018720819870700.
287. Wang, L., Li, Y., Lou, Y., Zhang, G., Chen, J., Wang, Y., & Zhang, X. (2017). Chemotherapy-related risk management toward safe administration of medications: Apply failure mode and effects analysis to reduce the incidence of chemotherapy errors. *Pakistan Journal of Pharmaceutical Sciences*, *30*(3).
288. WHO. (2002). *The importance of pharmacovigilance*. https://apps.who.int/iris/handle/10665/42493
289. WHO. (2017a). *Guidance on the Use of International Nonproprietary Names for Substances (INNs) for Pharmaceutical Substances*. http://apps.who.int/medicinedocs/documents/s23392en/s23392en.pdf
290. WHO. (2017b). *WHO Global Patient Safety Challenge: Medication Without Harm*. https://www.who.int/patientsafety/medication-safety/medication-without-harm-brochure/en/
291. WHO. (2020a). *List of Proposed and Recommended INNs*.
292. WHO. (2020b). *MedNet: INN search*. https://mednet-communities.net/inn/db/searchinn.aspx
293. WHO. (2020c). *WHO Forthcoming meetings*. World Health Organization. https://www.who.int/medicines/services/inn/meetings/en
294. WHO. (2021a). *ATC/DDD Index 2021*. https://www.whocc.no/atc_ddd_index/
295. WHO. (2021b). *INN Stem Book 2018*. https://www.who.int/publications/i/item/who-emp-rht-tsn-2018-1
296. WHO. (2023). *The International Pharmacopoeia*. WHO. https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/norms-and-standards-for-pharmaceuticals/international-pharmacopoeia

297. WHO. (2022). *International Nonproprietary Names (INN) for biological and biotechnological substances*. INN Programme. https://cdn.who.int/media/docs/default-source/international-nonproprietary-names-(inn)/inn-bio-review-2022.pdf?sfvrsn=f8db166f_5&download=true
298. WHO (2025) *Pre-stems: suffixes used in the selection of INN – October 2024.* https://www.who.int/publications/m/item/inn-24-584. Accessed 10/2/25.
299. WHO INN committee. (2016). *62nd Consultation on International Nonproprietary Names for Pharmaceutical Substances*. https://www.who.int/medicines/services/inn/62nd_Executive_Summary.pdf?ua=1
300. Wolak, E. (2017). A Novel Approach to Reducing RN Distraction During Medication Access. *Medsurg Nursing: Official Journal of the Academy of Medical-Surgical Nurses*, *26*(2).
301. Wong, Z. S. Y. (2016). Statistical classification of drug incidents due to look-alike sound-alike mix-ups. *Health Informatics Journal*, *22*(2), 276–292.
302. World Health Organization. (n.d.). *Revised monoclonal antibody (mAb) nomenclature scheme—Geneva 26 May, 2017*. Retrieved 2017, from https://cdn.who.int/media/ docs/default-source/international- nonproprietary-names-(inn)/revised-mab- nomenclature-scheme.pdf
303. Zafar, R., & Raza, M. L. (2020). Assessment of Look-Alike, Sound-Alike and Read-Alike (LASARA) Medicine--Errors in Pharmacy. *Journal of the College of Physicians and Surgeons--Pakistan: JCPSP*, *30*(4), 425–428.
304. Zaidi, S., Mordaunt, C., Durnin, N., Cooke, S., & Blakey, J. D. (2015). Quantifying and reducing inhaler prescription errors in secondary care. *International Journal of Clinical Pharmacy*, *37*(6), 1028–1032.
305. Zhong, W., Feinstein, J. A., Patel, N. S., Dai, D., & Feudtner, C. (2016). Tall man lettering and potential prescription errors: a time series analysis of 42 children's hospitals in the USA over 9 years. *BMJ Quality & Safety*, *25*(4), 233–240.
306. Zito, J. M., Safer, D. J., de Jong-van den Berg, L. T. W., Janhsen, K., Fegert, J. M., Gardner, J. F., Glaeske, G., & Valluri, S. C. (2008). A three-country comparison of psychotropic medication prevalence in youth. *Child and Adolescent Psychiatry and Mental Health*, *2*(1), 26. https://doi.org/10.1186/1753-2000-2-26.