Natural language processing (NLP) for clinical information extraction and healthcare research

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

Beata Fonferko-Shadrach
BSc MPH (989143)

Neurology Research Group
Swansea University Medical School
Swansea University

October 2022

Abstract

Introduction Epilepsy is a common disease with multiple comorbidities. Routinely collected health care data have been successfully used in epilepsy research, but they lack the level of detail needed for in-depth study of complex interactions between the aetiology, comorbidities, and treatment that affect patient outcomes. The aim of this work is to use natural language processing (NLP) technology to create detailed disease-specific datasets derived from the free text of clinic letters in order to enrich the information that is already available.

Method An NLP pipeline for the extraction of epilepsy clinical text (ExECT) was redeveloped to extract a wider range of variables. A gold standard annotation set for epilepsy clinic letters was created for the validation of the ExECT v2 output. A set of clinic letters from the Epi25 study was processed and the datasets produced were validated against Swansea Neurology Biobank records. A data linkage study investigating genetic influences on epilepsy outcomes using GP and hospital records was supplemented with the seizure frequency dataset produced by ExECT v2.

Results The validation of ExECT v2 produced overall precision, recall, and F1 score of 0.90, 0.86, and 0.88, respectively. A method of uploading, annotating, and linking genetic variant datasets within the SAIL databank was established. No significant differences in the genetic burden of rare and potentially damaging variants were observed between the individuals with vs without unscheduled admissions, and between individuals on monotherapy vs polytherapy. No significant difference was observed in the genetic burden between people who were seizure free for over a year and those who experienced at least one seizure a year.

Conclusion This work presents successful extraction of epilepsy clinical information and explores how this information can be used in epilepsy research. The approach taken in the development of ExECT v2, and the research linking the NLP outputs, routinely collected health care data, and genetics set the way for wider research.

Declaration

I, Beata Fonferko-Shadrach, confirm this thesis has not been submitted towards a previous degree or other qualification, and is intended for submission of a Doctor of Philosophy, awarded by Swansea University.

Signed: [redacted]
Date: 03/10/2022

I, Beata Fonferko-Shadrach, confirm that all of the work presented in this thesis is my own unless otherwise indicated. I have provided footnotes in the text where I have received assistance and have indicated permissions for presenting work which is not my own.

Signed: [redacted]
Date: 03/10/2022

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access approved by the Swansea University.

Signed: [redacted]
Date: 03/10/2022

Acknowledgements

I would very much like to thank my supervisors, Professor Mark Rees, for encouraging me to undertake this PhD, Professor Julian Halcox, for taking over the role at short notice, Dr Jonathan Mullins, and Dr Owen Pickrell for his support, guidance, and practical advice. I am very much indebted to him for introducing me to natural language processing, which started this research adventure. The work on this thesis would not be possible without his expertise and commitment.

Special acknowledgement and thanks must be given to my neurology data research colleagues, Dr Arron Lacey, for his work on the original NLP pipeline and his SQL expertise; Huw Strafford, for his work on the redeveloping ExECT and mastery of R, and to Carys Jones, who helped us all in the annotation tasks, and crucially Samuel Dobbie for creating Markup. None of this work would be possible without them.

I would like to thank the other members of the Swansea Neurology Research Group, past and present, who made this work possible and enjoyable: Dr Owain Howell, Dr Seokyung Chung, Dr Anna Derrick, Dr Adam Higgins, and Sarah Dawes.

I am very grateful to Mark Baker for his work on the Swansea Neurology Biobank sample and clinical data collection.

A special thank you must go to Professor Cathy White for her passion for epilepsy, and to the other members of the Morriston Hospital Neurology and Paediatric Neurology departments.

Finally, I must thank my family for being kind.

Papers and presentation 2019 – 2022

Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford D V., et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: Development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open. 2019 Apr 1;9(4).

Dobbie S, Strafford H, Pickrell WO, Fonferko-Shadrach B, Jones C, Akbari A, et al. Markup: A Web-Based Annotation Tool Powered by Active Learning. Front Digit Heal. 2021 Jul 26;3.

Beata Fonferko-Shadrach, Huw Strafford, Cathy White, Arron Lacey and William Owen Pickrell: Using natural language processing to extract features and results from free text electroencephalography (EEG) reports; Healthcare Text Analytics Conference HealTAC 2021, June 16-18, 2021, http://healtex.org/healtac-2021/

Fonferko-Shadrach B, Lacey AS, Strafford H, Jones C, Baker M1, Chung SK1, Jones KH Pickrell WO Rees MI Linking whole exome sequencing with electronic health care records – a proof of concept study. ILAE 14th European Epilepsy Congress, 9-13 July, Geneva, Switzerland

Contents

1 INTRODUCTION 2

1.1 Thesis structure . . . 2

1.2 Epilepsy . . . 3

1.2.1 Epilepsy and seizure classification . . . 4

1.3 Routinely collected healthcare data . . . 20

1.3.1 The SAIL Databank . . . 22

1.3.2 Epilepsy research using linked data . . . 23

1.4 NLP for Clinical Information extraction . . . 24

1.4.1 UMLS . . . 26

1.4.2 Gold standard annotation . . . 27

1.5 NLP for epilepsy . . . 30

1.6 Review of existing systems . . . 31

1.7 The genetics of outcomes in epilepsy . . . 33

1.7.1 Objectives . . . 35

2 MATERIALS AND METHODS 37

2.1 Document sourcing and preparation . . . 37

2.1.1 Development and test sets . . . 38

2.1.2 Gold standard annotation development and validation sets 38

2.1.3 Swansea Neurology Biobank database . . . 40

2.2 Gold Standard development . . . 42

2.2.1 Annotation guidelines for Epilepsy . . . 42

2.2.2 Inter-annotator agreement . . . 43

2.3 Annotation software . . . 45

2.3.1 BRAT . . . 45

2.3.2 Markup . . . 46

2.4 Algorithm development . . . 48

2.4.1 Tokeniser . . . 49

2.4.2 Named entities – Gazetteers . . . 49

2.4.3 ILAE classification . . . 51

2.4.4 Context implementation 51
2.4.5 Certainty levels 53
2.4.6 Rule development 54
2.4.7 Groovy script 56
2.4.8 ExECT v2 pipeline 57
2.5 Validation 58
2.6 Genetic data linkage project 59
2.6.1 Epi25 Collaborative 60
2.6.2 VCF files upload into SAIL 61
2.6.3 Annovar 61
2.6.4 Annovar within the SAIL Gateway 62
2.6.5 Annotating VCF files 62
2.6.6 Crating variant data table 63
2.6.7 Allele frequency 65
2.6.8 Combined Annotation-Dependent Depletion 65
2.6.9 Extracting variants 66
2.6.10 SAIL data 66
2.6.11 ExECT data 68
2.6.12 Linked data analysis 68
2.6.13 Project Approval 70
2.6.14 Version control 71
2.6.15 Literature search, library, and thesis writing 71
2.7 Chapter summary 71

3 GOLD STANDARD ANNOTATION 73
3.1 Standard annotation specification 73
3.1.1 Test and validation sets 74
3.1.2 Markup configuration 74
3.1.3 Annotation guidelines for epilepsy 75
3.2 Annotation tests 75
3.2.1 Measuring agreement - Fleiss' Kappa 76
3.2.2 Measuring agreement - pairwise F measure 78
3.3 Gold standard set for epilepsy 80
3.4 Chapter summary 81

4 ExECT v2 PIPELINE AND OUTPUT VALIDATION 83
4.1 ExECT v2 output 83
4.2 ExECT v2 Validation 85
4.2.1 Overall results – per item 85
4.2.2 Overall results – per letter 87

4.2.3 Results for specific entities 87
4.2.4 Diagnosis 89
4.3 Chapter summary 104

5 SUPPLEMENTARY NLP PIPELINE FOR EXTRACTING IDENTIFIABLE INFORMATION - IDEx 105
5.1 Extracting identifiable information from clinic letters 106
5.1.1 IDEx design and output 106
5.2 Validation 110
5.2.1 IDEx Validation on 200 epilepsy letters 110
5.2.2 Validation of the Epi25 cohort output against the SNB Donors database 111
5.2.3 Validation of Clinic, Letter, and Record Date extraction . . 116
5.3 Chapter summary 117

6 EPI25 COHORT PROCESSING - OUTPUT VALIDATION AND ANALYSIS 119
6.1 Per person Epi25 cohort dataset 120
6.1.1 The datasets 120
6.1.2 Documents processing 120
6.1.3 Pipeline outputs and data table construction 121
6.2 ExECT output 121
6.2.1 ExECT Epilepsy Diagnosis 121
6.2.2 Onset 127
6.2.3 Patient History 129
6.2.4 Investigations 132
6.3 Linking ExECT outputs for longitudinal analysis 137
6.3.1 Seizure Frequency 137
6.3.2 Linking seizure frequency and ASM outputs for the entire cohort 151
6.4 Chapter summary 157

7 NLP WITHIN SAIL DATABANK AND GENETIC DATA LINKAGE 159
7.1 Epilepsy genetics data linkage study 159
7.1.1 SNB cohort within the SAIL databank 160
7.1.2 Rare variants burden analysis 161
7.1.3 Exome-wide burden 162
7.1.4 Gene-based variant burden 163
7.1.5 Genetic burden and Unscheduled Hospital Admissions . . . 163
7.1.6 Genetic burden and GP records of ASM 165

7.1.7 Genetic burden analysis by epilepsy type 166
7.2 Linking NLP output to the genetic study within SAIL gateway . . 170
7.2.1 Per person Epi25 cohort NLP pipeline outputs 170
7.2.2 Seizure frequency ExECT output linkage 171
7.3 Chapter summary 176

8 DISCUSSION 177
8.1 Gold standard annotation 177
8.1.1 Annotation guidelines 177
8.1.2 Documents 179
8.1.3 Annotation tasks 180
8.2 ExECT development and validation 183
8.2.1 Onset and When diagnosed 187
8.2.2 Epilepsy cause 188
8.2.3 Certainty 189
8.2.4 Strengths and limitations 190
8.3 IDEx 191
8.4 Epi25 cohort processing – output validation and analysis 192
8.4.1 The SNB and the Epi25 cohort 193
8.4.2 ExECT validation on the Epi25 SNB cohort 193
8.4.3 Linking ExECT outputs for longitudinal analysis 196
8.5 Genetic linkage study 199
8.5.1 Creating a pipeline for genetic data linkage 199
8.5.2 Genetic burden, unscheduled admissions and ASM records . 200
8.5.3 Genetic burden analysis by epilepsy type 202
8.5.4 NLP output linkage within the SAIL databank 203
8.5.5 Seizure frequency and genetic burden 203
8.6 Conclusions 205

A The Swansea Neurology Biobank 208
A.1 Consent 208
A.2 Biobank database 208

B Gold standard annotation set 213
B.1 Markup configuration for epilepsy clinic letter annotation 213
B.2 What and how of annotating with Markup 213

C Epi25 Cohort processing – output validation and analysis 233
C.1 Diagnosis in the SNB database 233
C.2 Epi25 Onset processing 233
C.3 Seizure frequency processing 233
C.4 Prescription processing (fragment) 233

D NLP within SAIL Databank and Genetic Data Linkage 242
D.1 Linking Epi25 cohort with the SAIL databank datasets 242
D.2 Epilepsy and drug metabolism and transportation genes 242
D.3 Filtering for rare and potentially damaging variants 242
D.4 Residual Variation Intolerance Scores for genes affected by variants identified in the linkage 242

List of Figures

2.1 Document preparation scripts 39
2.2 Annotation Guidelines Development 43
2.3 Markup annotation tool 47
2.4 Markup annotation output 48
2.5 UMLS based gazetteers 52
2.6 ConText and Certainty ExECTv2 annotation in GATE 54
2.7 JAPE rule for Onset 55
2.8 Groovy script for Epilepsy cause 57
2.9 ExECT v2 pipeline 57
2.10 Genetic data linkage pipeline - uploading VCF files, clinical data from the SNB, ExECT v2 output, and individual demographic data for linkage with the routinely collected data within the SAIL databank 60
2.11 VCF in .avinput text file structure 63
2.12 Perl script creating a single VCF file 64
2.13 VCF file following annotation 65
2.14 SAIL genetics data linkage pipeline 70

4.1 Groovy script for extracting Patient History annotations 84

5.1 JAPE rule for Hospital number annotation 107
5.2 JAPE for letter date annotation 108
5.3 Groovy script for Date of Birth annotation output and the list of features extracted 112
5.4 Validating IDEx against the SNB Donors database using sqldf in R studio. Script shows NHS number, Gender, and Postcode validation 113

6.1 Epi25 cohort ExECT output – diagnosis filtering 123
6.2 Epi25 cohort ExECT output for Age of onset after before processing 128
6.3 Epi25 cohort ExECT output for Age of onset after processing . . . 129

6.4 Seizure Frequency output processing flow diagram 139
6.5 Seizure Severity Score calculation 139
6.6 Seizure severity score over a period of time, Person One 140
6.7 Seizure Severity Score over a period of time, Person Two 141
6.8 ExECT v2 seizure frequency annotation within the GATE developer 142
6.9 Seizure Severity Score over a period of time, Person Three 143
6.10 Prescription output processing flow diagram 145
6.11 ASM and seizure frequency, Person One 147
6.12 ASM and seizure frequency, Person Two 148
6.13 ASM and seizure frequency, Person Three 150
6.14 ASM monotherapy and polytherapy and seizure frequency, all seizures 152
6.15 ASM monotherapy and polytherapy and seizure frequency, GTCS . 153
6.16 ASM monotherapy and seizure frequency, by ASM 154
6.17 ASM polytherapy and seizure frequency, by ASM combinations . . 155
6.18 ASM polytherapy and seizure frequency, by ASM combinations of 3 or 4 drugs 156

7.1 Summary of the variant filtering process 161
7.2 SQL script creating Cumulative Burden table 162
7.3 Genetic variants and unscheduled hospital admissions 164
7.4 Gene, variants, and unscheduled hospital admissions – Venn diagram 165
7.5 Genetic variants and ASM prescriptions 166
7.6 Gene, variants, and ASM prescriptions – Venn diagram 167
7.7 CADD score and number of qualifying variants by epilepsy type . 168
7.8 Epilepsy and Drug genes and variants by epilepsy type 169
7.9 Seizure frequency linkage 172
7.10 Seizure frequency and genetic burden 174
7.11 Seizure frequency and genes associated with epilepsy 175

A.1 Figure A1.1: SNB Consent form, page 1 209
A.2 Figure A1.1: SNB Consent form, page 2 210
A.3 Figure A1.2: SNB clinical database epilepsy contents 211
A.4 Figure A1.3: SNB Donors database, Participant form 212

B.1 Markup configuration page 1 214
B.1 Markup configuration page 2 215
B.1 Markup configuration page 3 216
B.2 Annotation guidelines page 1 217
B.2 Annotation guidelines page 2 218
B.2 Annotation guidelines page 3 219
B.2 Annotation guidelines page 4 220

B.2 Annotation guidelines page 5 221
B.2 Annotation guidelines page 6 222
B.2 Annotation guidelines page 7 223
B.2 Annotation guidelines page 8 224
B.2 Annotation guidelines page 9 225
B.2 Annotation guidelines page 10 226
B.2 Annotation guidelines page 11 227
B.2 Annotation guidelines page 12 228
B.2 Annotation guidelines page 13 229
B.2 Annotation guidelines page 14 230
B.2 Annotation guidelines page 15 231
B.2 Annotation guidelines page 16 232

C.1 r script creating single epilepsy diagnosis table from from the SNB database record of epilepsy and seizure using ExECT list od epilepsy and seizure terms 234
C.2 R script converting `time since', 'date when', and 'age'into a single common measure of age of onset. 235
C.2 R script converting `time since', 'date when', and 'age'into a single common measure of age of onset. 236
C.3 R script processing seizure frequency output from ExECT into a dated record of seizure scores. 237
C.3 R script processing seizure frequency output from ExECT into a dated record of seizure scores. 238
C.3 R script processing seizure frequency output from ExECT into a dated record of seizure scores. 239
C.4 R script processing prescription output from ExECT into a dated record of the total daily dose of ASM. 240
C.5 R script processing prescription output from ExECT into a dated record of the total daily dose of ASM. 241

D.1 Epi25 cohort linkage within the SAIL databank 243
D.2 Genes associated with epilepsy, drug metabolism and transportation used in the gene-based analysis. Source:**A** Epi25 Collaboraive, epilepsy-genes, [1]; **B** Epilepsy associated genes, [2]; **C** Drug metabolising enzymes and transporters. [3] 244
D.3 SQL script filtering rare and potentially damaging variants from the annotated whole exome sequenced data table for the study cohort 245

D.4 RVIS for genes identified following filtering for rare and potentially damaging variants and not occurring more than twice in the study cohort in different groups of individuals defined by unscheduled hospital admissions (No admissions versus Admissions), ASM (monotherapy versus polytherapy), and seizure frequency (seizure freedom for > 1 year versus at least 1 seizure per year. Epilepsy associated genes (black font), drug metabolism and transportation (green font)) 246

List of Tables

1.1 ILEA 2017 classification of focal seizures. 6
1.2 ILEA 2017 classification of either focal or generalised onset seizures. 7
1.3 ILEA 2017 classification of generalised onset seizures. 7
1.4 Genes in common epilepsies 13

2.1 SNB – extract of diagnosis and seizure information 41
2.2 Seizure onset information in the SNB 42
2.3 Confusion matrix for the evaluation of information extraction . . . 58

3.1 Fleiss' Kappa for IAA 77
3.2 Pairwise average F1 score for IAA 79

4.1 Entities extracted by ExECT 85
4.2 ExECT v2 'per item' validation, against the gold standard set . . . 86
4.3 ExECT v2 'per letter' validation against the gold standard set . . 88
4.4 ExECT v2 validation of epilepsy diagnosis by type 91
4.5 *ExECT v2 validation of CT, MRI, and EEG results against gold standard set, per item and per letter scores* 93
4.6 ExECT v2 Patient History annotations 96

5.1 IDEx entities and feature description 109
5.2 IDEx validation on the 200-letter set 110
5.3 Epi25 cohort IDEx validation against the SNB – per letter 114
5.4 Epi25 cohort IDEx validation against the SNB – per person 114
5.5 Validation of Clinic, Letter, and Record dates on a random sample from the Epi25 letter-set 117

6.1 Epi25 ExECT output for epilepsy diagnosis validation against the SNB records 124
6.2 Epi25 cohort analysis of 'missing'epilepsy diagnoses 125

6.3 Epi25 cohort ExECT output for Age of onset vs the SNB records error review 130
6.4 Epi25 cohort ExECT output for ‘no history of febrile seizures’validation against the SNB 131
6.5 Epi25 cohort ExECT output for EEG results by test type 133
6.6 The SNB EEG results records by test type 133
6.7 Epi25 cohort ExECT EEG results output comparison against the SNB database 134
6.8 Epi25 cohort the SNB database MRI records 135
6.9 Epi25 cohort MRI comparison against the SNB database 135

7.1 SNB cohort group characteristics 160
7.2 ExECT output groupings for the SAIL upload 170

Abbreviations

AD Alzheimer's disease

ADHD attention-deficit hyperactivity disorder

ALF anonymous linking field

ASD autism spectrum disorder

ASM antiseizure medication

BNF British National Formulary

CAE childhood absence epilepsy

CBZ carbamazepine

CLEF Clinical E-Science Framework

CNS central nervous system

CP cerebral palsy

CRF conditional random field

CUI Concept nique Identifier

EE epileptic encephalopathies

EEG electroencephalography

EHR electronic health records

EMAtS epilepsy with myoclonic atonic seizures

ESL eslicarbazepine

GGE genetic generalised epilepsy

GPRD General Practice Research Database

GTCA generalized tonic-clonic seizures alone

GTCS generalised tonic-clonic seizures

HDR Health Data Research

HITEX Health Information Text Extraction

i2b2 Informatics for Integrating Biology & the Bedside

IAA inter annotator agreement

ICD International Classification of Diseases

IGE idiopathic generalised epilepsy

ILAE International League Against Epilepsy

JAE juvenile absence epilepsy

JME juvenile myoclonic epilepsy

LCM lacosamide

LEV levetiracetam

LGS Lennox-Gastaut Syndrome

LMT lamotrigine

MedLEE Medical Language Extraction and Encoding System

MRI magnetic resonance imaging

MS multiple sclerosis

MTLE mesial temporal lobe epilepsy

n2c2 National NLP Clinical Challenges

NCBC National Centre for Biomedical Computing

NER named entity recognition

NIH National Institute of Health

NLP natural language processing

PPV positive predictive value

PREF Preferred Term

PRS polygenic risk scores

PST part of speech tagger

RALF Residential Anonymous Linking Field

SAIL Secure Anonymised Information Linkage

SANAD Standard and New Antiepileptic Drugs

SBUHB Swansea Bay University Health Board

SNB Swansea Neurology Biobank

SNV copy number variant

SSS seizure severity scores

STY Full Name Semantic Type

SVM support vector machines

TLE temporal lobe epilepsy

TPM topiramate

TTP trusted third party

TUI Type Unique Identifier

UIMA Unstructured Information Management Architecture

UMLS Unified Medical Language System

VPA sodium valproate

WES whole exome sequencing

WIMD Welsh index of multiple deprivation

ZNS zonisamide

Chapter 1

INTRODUCTION

This chapter describes the structure of the thesis. It presents the background and the main themes of the development of clinical Natural Language Processing which are then discussed in the context of epilepsy.

1.1 Thesis structure

This thesis describes the work on developing a Natural Language Processing pipeline for extracting information from clinic letters, including the processing of the structured outputs and examples of how these outputs could be used in research. Epilepsy is used as an example disease as it is a common neurological disorder, yet the detailed diagnostic and outcomes' information is not available in routinely collected data.

The overall aim of the project was the development of a system that could convert free text information contained in clinic documents, into structured output that could be analysed and linked to routinely collected data to enrich it for research.

As the development is applied to epilepsy and the linkage of the extracted data is used in an epilepsy genetic study, an introduction to the disease itself and to the main issues affecting it is made.

Chapter 2 presents the materials and methods used in this research. Chapter 3 describes the development of a gold standard annotation set for epilepsy clinic letters. Chapter 4 gives the details of the ExECT v2 outputs and the results of the validation process. Chapter 5 presents a pipeline for the extraction of personal demographic information, and gives the results from a number of validation tests. Chapter 6 presents the outputs of processing clinic documents for a cohort of individuals from the Epi25 study, and explores how these outputs could be linked in research. Chapter 7 describes a genetic study which linked exome sequencing data with routinely collected information and the output from the ExECT v2 pipeline within the SAIL databank. The final chapter discusses the results and possibilities for future research.

1.2 Epilepsy

The development of the NLP pipeline and the genetic linkage study described in this thesis relate to epilepsy. This section provides an introduction to the disease, describing its presentation, classification, epidemiology, treatment, comorbidities, and impact.

Epilepsy has been defined as a '*disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiologic, cognitive, psychological, and social consequences of this condition.*' [4] With an epileptic seizure being '*a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain*'. [5] In practical terms, this has meant that having two unprovoked seizures more than 24 hours apart was considered as diagnostic of epilepsy. This interpretation was amended by the International League Against Epilepsy (ILAE) in 2014 with the following criteria:

(i) '*At least two unprovoked (or reflex) seizures occurring more 24 hours apart*';

(ii) '*One unprovoked (or reflex) seizure and a probability of further seizures* *similar to the general recurrence risk (at least 60%) after two unprovoked seizures, occurring over the next 10 years*';

(iii) '*diagnosis of an epilepsy syndrome.*' [6]

1.2.1 Epilepsy and seizure classification

The classification of seizures[1] into focal (partial) or generalised was recommended by the ILAE in 1964, based on a clear distinction in the onset location. [7] This dichotomy remained, with some modifications, until 2017 when a new operational classification was introduced. The reason behind this change was to clarify nomenclature, allow for classification of some seizure types as either focal or generalized, and to make it possible to classify seizures as unknown onset. [8]

The mode of onset still remains a distinctive feature, with focal seizures defined as originating within neuronal networks limited to one hemisphere, localised or more widely distributed. Generalised seizures originate within, and rapidly engage bilaterally distributed neuronal networks. Individual seizure onsets may appear localized, but location and lateralization are not consistent from one seizure to another. [9] Unknown onset refers to seizures when the onset is unknown but other manifestations are known. [8]

During a focal seizure a person may be aware of self and the surrounding environment, or they may have an impaired awareness. Focal aware seizures (previously called simple partial seizures) may be motor or non-motor. Focal impaired awareness seizures (previously known as complex partial or focal dyscognitive seizures) means that during any part of the seizure a person is not aware of self or their environment. A focal seizure may spread and involve both hemispheres, with impaired consciousness and tonic-clonic elements, this is referred to as a focal to bilateral tonic-clonic seizure (previously termed secondary generalised tonic-clonic

[1] Seizure is used here to mean an epileptic seizure and when other seizures are mentioned this is clearly stated

seizure). [10] Focal seizures are the most common type of seizures, reported in about 60% of individuals with epilepsy. [11]

Generalised onset seizures may be motor, such as myoclonic or tonic-clonic, or non-motor i.e., absences, which are a sudden short staring episodes, table 1.3. Absence seizures are more common in children under the age of 15 and are associated with two distinct and common syndromes, childhood absence epilepsy (CAE), juvenile absence epilepsy (JAE). [12]

Tables 1.1 to 1.3 provide descriptions of seizures grouped by onset type according to the 2017 ILAE classification. [8] Table 1.1 lists focal onset seizures, table 1.2 seizures that can be of either focal or generalised onset, and table 1.3 generalised onset seizures. Tonic-clonic seizure which is listed here may also be classified as of unknown onset if the onset is not known, as may also apply to epileptic spasms and behavioural arrest. It may not be possible to categorise some seizures, due to the insufficient information available, and these should be classed as unclassified. [8] The format of the lists was intended to highlight the change from the previous classification as it is relevant to the concept extraction algorithms that are discussed in this thesis.

Focal onset seizures	
Motor Onset	
Automatism	Coordinated, repetitive motor activity. May involve lip smacking, swallowing, or blinking (orofacial); pedal, vocal, verbal, or sexual behaviours.
Paresis/paralysis	Weakness or complete paralysis of a muscle or group of muscles.
Hyperkinetic seizure	Involves irregular large amplitude movements, such as pedalling, pelvic thrusting, jumping, or thrashing.
Hemiclonic seizure	Sustained rhythmic jerking involving one side of the body at seizure onset.
Non-motor Onset	
Sensory	Sensation being experienced at seizure onset, without clinical signs of a seizure being observed. for example may relate to visual. auditory, olfactory, or gustatory sensations.
Cognitive	Alteration of cognitive function, a deficit or increase. For example expressive or receptive dysphasia/aphasia, dyslexia, déjà vu, dissociation, or left-right confusion.
Automatic	Relating to functions controlled by the automatic nervous system„ e,g., palpitations, flushing, or altered respiration.
Emotional	Affecting mood or emotion, e.g., panic, laughing, crying, anger.

Table 1.1: *2017 ILAE operational classification of seizures: focal onset seizures by feature.*

Besides the presentation features, focal seizures may be categorised by their anatomical location as:

- Frontal lobe seizures — brief involving a wide range of motor features. May begin with an aura, involve vocalisation, odd movements, and head deviation. They occur often in sleep.
- Temporal lobe seizures — behavioural arrest, automatisms and sensory (fear, déjà vu, epigastric sensations) are common. An important subtype are mesial temporal lobe seizures with a range of automatic, cognitive, and sensory features. They may progress to impaired awareness and automatisms.
- parietal lobe seizures — sensory disturbance, disorientation, visual hallucinations, and language disturbance may occur.
- Occipital — focal sensory visual seizures, eyelid fluttering and eye deviation may occur.

Generalised or focal onset	
Atonic	Previously only with generalised onset. Sudden loss of muscle tone, falling, short-lasting with rapid recovery.
Clonic seizure	Sustained rhythmic jerking to involve parts of the body according to their representation on the motor cortex (known as a Jacksonian march).
Epileptic spasms	a sudden flexion, extension or mixed flexion-extension of proximal and truncal muscles, lasting 1-2 seconds and usually occurring in a series.
Myoclonic	Very brief muscle contractions (jerks), occuring singularly or in clusters.
Tonic	Increased muscle tone, usually lasting for seconds to minutes, in generalised onset a person is unaware during the attacks

Table 1.2: *2017 ILAE operational classification of seizure type: seizures that could have either focal or generalised onset, by features.*

Generalised onset seizures	
Motor Onset	
myoclonic-atonic	Myoclonic seizure followed by an atonic seizure, resulting in rapid fall.
tonic-clonic	Bilateral and symmetric generalized motor seizures with loss of consciousness. It consists of a tonic (bilateral increased tone) and then a clonic (bilateral sustained rhythmic jerking) phase. Followed by postictal confusion. May be classified as unknown onset if the onset is unclear.
Non-motor (Absence)	
Typical absence	Sudden, abrupt onset and offset of a staring episodes with rapid recovery. Clonic movements of eyelids, head, eyebrows, chin may occur
Atypical absence	Similar to a typical absence with less abrupt onset and offset. Often associated with other features such as loss of muscle tone and myoclonic jerks. The loss of awareness may be minimal. Associated with intellectual impairment.
Myoclonic absence	Absence with myoclonic jerks of the shoulders and arms, typically bilateral.
Absence with eyelid myoclonia	Short absence seizure with brief, repetitive, often rhythmic, myoclonic jerks of the eyelids with simultaneous upward deviation of the eyeballs and extension of the head.

Table 1.3: *2017 ILAE operational classification of seizurs: generalised onset seizures, by features.*

In 2017 the ILAE produced and updated classification of epilepsies which brought a significant change from the 1989 version, by the addition of a distinct category of ‘combined generalised and focal epilepsy’ to the established focal and generalised epilepsies. [13] The ‘Unknown’ category is used for cases where it is clear that an individual has epilepsy, but it is not possible to classify it under the other three types.

Focal epilepsies are characterised by focal onset seizures and can be unifocal and multifocal. The diagnosis is based on clinical presentation but may be supported by the electroencephalography (EEG) showing focal epileptiform discharges. [14] Magnetic resonance imaging (MRI) can identify structural lesions, although in up to a third of individuals there are no obvious epileptogenic lesions. [15]. The disease classification into specific focal epilepsy subgroup or a syndrome is based on aetiology and age. For example childhood epilepsy with centrotemporal spikes (Rolandic epilepsy) is a self-limiting syndrome with brief, hemifacial seizures which may evolve to tonic-clonic seizures, occurring between the ages of 4–14 years in the context of normal development. Benign occipital epilepsy such as Panayiotopoulos syndrome is characterised by autonomic symptoms (mainly vomiting) affecting children aged 3–6 years. [16]
In temporal lobe epilepsy (TLE), the age of onset may be associated with seizure semiology, [17] but in general, the outcome is more related to the disease aetiology, with the lesional caused being associated with intractable seizures. [18]

Generalised epilepsies are diagnosed on the evidence of generalised onset seizures e.g., absences, myoclonic or tonic-clonic seizures, and topically supported by generalized spike-wave activity on EEG. [14] A significant subgroup of generalised epilepsies are genetic generalised epilepsies (GGE), where seizures are associated with a clear generalised spike-wave EEG patterns. Within these, childhood absence epilepsy (CAE), juvenile absence epilepsy (JAE), juvenile myoclonic epilepsy (JME), and generalized tonic-clonic seizures alone (GTCA), are a distinct group with a defined set of seizures and the EEG pattern with 2.5–5.5Hz spike-wave, and a polygenic inheritance. This group is referred to as idiopathic generalised epilepsies (IGEs) and it accounts for some 15%–20% of all GGEs. [19]

Other GEE include epilepsy with eyelid myoclonia, which is a rare disease with onset at 6–8 years, presenting with eyelid myoclonia and generalised tonic-clonic seizures (GTCS), and epilepsy with myoclonic absence, with onset at 1–12 years. [20] Whereas these syndromes together CAE have a specified childhood onset, other GEE may have a variable onset age.

Developmental epileptic encephalopathies (DEE) are a group of syndromes in which the epileptic activity contributes to developmental slowing and regression. Epileptic encephalopathy (EE) describes an assumed causal relationship between epilepsy and developmental delay, whereas developmental encephalopathies (DE) are more independent from epilepsy i.e., they continue even when seizures are controlled. [21] Some of these are generalised epilepsies such as epilepsy with myoclonic atonic seizures (EMAtS), with the onset in the early childhood and, in most cases, in the context of normal development. [20]

Lennox-Gastaut syndrome (LGS) and Dravet syndrome are two examples of DEE which fall into the category of combined generalized and focal epilepsies as patients have both generalised and focal seizure types. [13]

Epidemiology

A recent review of population-based studies of prevalence and incidence of epilepsy worldwide, reported point prevalence of active epilepsy at 6.38 per 1,000 individuals (95% CI 5.57–7.30) with slightly higher rates reported for low-middle income countries. [22] Similar rates were observed in a number of European population studies, with the rates ranging from 3.4 to 7.8 per 1,000 individuals. [23] The median incidence rate for epilepsy based on a meta-analysis, was 69 per 100,000 per year in the developing countries and 43 per 100,000 in the industrialised countries. [24] In Wales an estimated mean prevalence of epilepsy was reported as 0.77% (95% CI 0.76 – 0.79%), and the estimated incidence rate as 29.5 per 100,000 per year (95% CI 28.3 to 30.7). [25]

The increased incidence rates in developing countries could be due to a combination of factors, different population structure, greater risk of central nervous system (CNS) infections, or brain injuries, but also methodological issues in case verification. [26]

Socioeconomic factors play a significant role in prevalence of epilepsy. In developing countries, a marked difference has been reported in the rates between rural and urban populations, 15.4 per 1,000 as compared to 10.3 per 1,000 respectively. [27] There are variations in the range of indicators used in studies investigating the link between deprivation and epilepsy, however, the measures used aim to identify, in one way or another, individuals and communities that are disadvantaged in terms of resources and opportunities. Using this definition, a number of studies have reported an association between the incidence and prevalence of epilepsy and deprivation. [28–31] One explanation for this has been sought in the social drift hypothesis of downward social mobility due to impaired health, mental disorders, and reduced income. But there could also be an association between epilepsy risk factors that may be more prevalent in deprived communities, for example trauma, infection, or poor nutrition [32]

A linkage study of epilepsy prevalence and incidence and deprivation in Wales used the Welsh index of multiple deprivation (WIMD) and primary care records. Epilepsy prevalence reported ranged from 1.13% (1.07–1.19%) in the most deprived and 0.49% (0.45–0.53%) in the least deprived decile. Epilepsy incidence rate showed a similar correlation, with 40/100,000 per year in the most deprived areas and 19/ 100,000 per year in the least deprived, suggesting that a causative link between deprivation and epilepsy. [25] Similar findings were reported by a study of the annual incidence of first unprovoked seizures and new diagnosis of epilepsy in Cork, Ireland. [33]

The incidence of epilepsy varies with age, with higher rates observed in the youngest (under 1 year) and oldest age groups, and with the highest rate in people over 75 years. [34] This pattern may be explained by the association of perinatal events, birth trauma, or congenital malformations with the early onset of epilepsy and by the increased risk of epilepsy due to the increased rates of cerebrovascular disease, degenerative disorders, intracerebral tumours in the older population. [35, 36]

Focal epilepsy affects about 60% of people with epilepsy as compared to generalised, [37] although the incidence varies by age, as generalised epilepsy is more frequent in young people, for under 15 year-old it is 55% as compared to 46% for focal epilepsy. [38]

Causes of epilepsy

Epilepsy is a complex disease with many underlying causes, and comorbidities that play a part in the course of the disease. Some distinction can be made between epilepsies with a recognised cause, such as brain injury or neurodegenerative disease, and genetic causes, for which there is no underlying causal factor, although this division is not clear cut. The 2017 ILAE classification of epilepsies divided epilepsy aetiologies into six categories: genetic, structural, metabolic, infectious, immune and unknown. The groupings are not exclusive, for example certain malformations of cortical development (structural causes) are genetic in origin, for instance, polymicrogyria or tuberous sclerosis. Similarly, some of the metabolic causes such as glucose transporter deficiency are due to genetic factors, i.e.*SLC2A1* mutations. Some examples of the main causes are listed here with a more detail discussion of genetics in relation to epilepsy outcomes later in this introductory chapter, as it is relevant to the linkage study presented in this thesis.

The genetic aetiology of epilepsy means that there is an established causal association between known or assumed genetic mutation and seizures. These mutations may relate to: missense variants which produce a substitution of one amino acid for another and can cause loss-of-function, gain-of-function, or can be functionally benign; truncating variants resulting in shortened protein that is usually non-functional; indels which are small insertions or deletions of one or a few nucleotides; and copy number variants (CNVs) which are large insertions or deletions, sometimes involving thousands of nucleotides. [39]

Most of the epilepsy genetic discoveries have been made in relation to rare and/or severe monogenetic forms of the disease.
For some of these the association is based on monogenic familial inheritance. For example mutations of potassium channel genes *KCNQ2* and *KCNQ3* have

been identified in benign familial neonatal epilepsy (BFNE). [40] In familial focal epilepsy with variable foci (FFEVF), family members have seizures originating in different brain regions, *DEPDC5* mutations have been identified as a common cause. [41]
Monogenic epilepsies can arise from *de novo* mutation, in Dravet syndrome, for example, more than 75% of individuals affected have a pathogenic variant of *SCN1A*, of these 90% are *de novo* . [42] Dravet syndrome is a form of early onset epilepsy with multiple seizure types and developmental delay. There are many other epilepsies associated with *SCN1A* mutations, ranging from severe DEEs to milder diseases such as genetic epilepsy with febrile seizures plus (GEFS+). [43]
Most of the DEEs arise from *de novo* mutations and there are some that are monogenic, although there may be other genes involved in a number of cases. For example, *KCNT1* is associated with more than 50% of cases of epilepsy of infancy with migrating focal seizures, but *SCN2A*, and *SCN2A* may be involved in up to 50% of cases. In early-onset epileptic encephalopathy *KCNQ2* is associated with up to 50% of cases, but there are a number of other genes implicated. Many DEEs are polygenic, meaning that there is no gene of major effect but rather an interaction between several genetic variants. [44]

In GGE which is the most common form of genetic epilepsy, the evidence of genetic bases is derived from studies of population with the same syndrome. Twin studies have shown that the recurrence risk of common GGE syndromes in monozygotic twins is higher than in dizygotic twins. [45] But the rate of epilepsy in siblings of individuals with epilepsy is lower that could be expected in dominantly inherited traits, suggesting that in most cases GGE is a result of multiple gene variants. [46] In this and in other common epilepsies single gene association discoveries have been limited to a small number of cases.

Advances in technology and the availability of larger number of genetic samples through collaborative efforts such as EpiK or EPIGEN, and more recently Epi25, have facilitated the discovery of common and rare gene variants associated with common epilepsies. These epilepsies comprise some 95% of all genetic generalised and focal epilepsies. Some of the most important genes implicated are shown in table 1.4

GGE	
CAE	calcium channel genes: *CACNA1H, CACNG3*, Acetylcholine receptor *CHRNA4* , Glutamate receptor *GRM4*
	GABA A and B receptor genes: *GABRG2, GABRA1, GABRB3, GABAB1, GABAB2*; glucoe transporter GLUT1: *SLC2A1*
JME	*GABRA1, EFHC1, CLCN2* and potassium ion channels: *KCNQ2, KCNQ3*, Nucleic acid binding: *BRD2*
JAE	*CACNB4*, *GABRA1, GRIK1* and *EFHC1*
Focal epilepssy	
TLE	*SCN1A*, *CALHM1*

Table 1.4: *Genes implicated in common epilepsies. [47]*

Studies have shown that some rare variants of epilepsy genes (*KCNQ2, SCN1A, and GABRG2*) are present in common epilepsies. [48] The association of ultra-rare coding variation with common epilepsies was also shown in the Epi25 whole exome sequencing (WES) results, from which studies reported a significant mutational burden in DEE and GEE, confirming shared genetic characteristics between these diverse types of epilepsies. [49] An important example is *SLC2A1*, where loss of function mutations may result in GLUT1 deficiency leading to metabolic encephalopathy including intractable epilepsy, complex motor dysfunction, and intellectual disability. GLUT1 deficiency is seen in 10% of children with typical absence seizures starting before 4 years of age and around 1% of people with typical GGE overall. [46]
This exemplifies not only a varied presentation of a genetic variant effect but also the overlap that exists between the genetic causes of epilepsy and other diseases.

Structural epilepsies are characterised by the presence of distinct brain abnormality that may be acquired, i.e. brain trauma, stroke, or have a genetic origin, such as developmental abnormalities. The qualifying factor is that the changes are confirmed through brain imaging.
Two relatively common types of cortical development malformations are focal cortical dysplasia, which is associated with focal seizures, and tuberous sclerosis, which

is associated with early onset epilepsy with epileptic spasms and focal seizures.
Hippocampal sclerosis has a strong association with epilepsy and particularly with mesial temporal lobe epilepsy (MTLE) where it has been reported in over 50% of cases. [50]
Cerebrovascular disease is among the most common causes of epilepsy in adults [11] with the main risk factors identified as cortical involvement, haemorrhage, and early seizures. [51] Traumatic brain injury is also a common cause of epilepsy, accounting for about 20% of symptomatic cases of the disease. [52]
Brain tumour-related epilepsy is common in gliomas, with the risk of seizure varying from 60% to 100% among low-grade gliomas and from 40% to 60% in high-grade. Following surgery some 15% – 35% of individuals may have intractable seizures. [53]

Metabolic disorders can be associated with epilepsy, where seizures are one of the presenting symptoms, although they may be the main manifestation. [54] Seizures may be caused by different mechanisms, for example by accumulation of ammonia due to urea cycle disorders that may lead to brain damage, creatinine disorders associated with GTCS, or glucose transporter type 1 (GLUT-1) deficiency also associated GTCS, early onset absence seizures, but also focal seizures. [54]

Central nervous system (CNS) infections are associated with unprovoked seizures with reported risk between 6.8% to 8.3%. [55] The risk of developing epilepsy in the long term is greater in individuals who experience early (provoked) seizures during/following the infection, with reported 22% risk for viral encephalitis and 13% for bacterial meningitis. [56]

Immune epilepsies arise directly as a result of an autoimmune response and mostly consist of autoimmune limbic encephalitis, with inflammation of the limbic area causing seizure, memory loss, unconsciousness, and psychiatric symptoms. [57]

Comorbidity and mortality

Comorbidities impose a significant burden on people with epilepsy. In a recent 10-year follow up study of about 1000 individuals with epilepsy, 26% had at least one comorbidity. [58] With the most common being developmental/perinatal (7.5% of cases), psychiatric (6.2%), cardiovascular (5.3%), and endocrine/metabolic (3.8%).

The relationship between another disease and epilepsy can be classified as direct causative association such as exists in the context of cardiovascular diseases, where in over 10% of individuals it will cause epilepsy. [59] Another example is multiple sclerosis (MS) where the age-adjusted prevalence is about 3-fold higher than that of the general population. [60]

A reverse of such causality is the case of fractures following seizures. For example, in a study of admissions with a diagnosis of seizure , 1.1% of 2,800 individuals sustained fracture of which 0.3% were direct consequence of seizure. [61]

Another type of association is shared risk when epilepsy and another disease have common risk factors. These risks are mainly genetic, where both epilepsy and comorbidity arise from genetic causes, for example *SCN1A* mutations cause Dravet syndrome, GEFS+ and other epilepsy syndromes, but it is also associated with hemiplegic migraine, autism spectrum disorder (ASD), and sudden death. [62] The association between type 1 diabetes and epilepsy might be related to the shared presence of anti-glutamic acid decarboxylase (GAD) antibodies, which are strongly associated with type 1 diabetes (in about 80% of individuals) and are present in up to 6% of people with epilepsy. A retrospective 25-year follow up study of newly diagnosed individuals with type 1 diabetes reported that they were 3 times as likely to develop epilepsy as compared with matched controls without type 1 diabetes. [63]

Other studies report similar results but also a number of associations that are more difficult to explain causally. For example, the Canadian community health survey reported that individuals with epilepsy were more likely to report lifetime anxiety disorders or suicidal thoughts with odds ratios of 2.4 and 2.2 respectively. [64] The same survey also reported that people with epilepsy had a statistically significantly higher prevalence of most chronic conditions than the general population. Conditions with particularly high prevalence in epilepsy (prevalence ratio ≥ 2.0) included stomach/intestinal ulcers, stroke, urinary incontinence, bowel disorders, migraine, Alzheimer's disease, and chronic fatigue. [65] The risk of epilepsy in Alzheimer's disease (AD) is higher than in other age-matched controls without dementia and is further elevated in early-onset of familial AD. [66] Individuals with mutations in three known genes for (AD) (*PSEN1, PSEN2, APP*)show a dramatic elevation of epilepsy risk. [67] Autism spectrum disorder (ASD) includes epilepsy as a feature in up to one-third of individuals. [43] Depression is the most common psychiatric comorbidity of epilepsy with lifetime prevalence of 35% . [64]

Dissociative seizures (also known as non-epileptic seizures, non-epileptic attacks, psychogenic seizures and pseudoseizures) resemble epileptic seizures, but have no EEG correlate or clinical evidence for epilepsy. It is reported that some 10% of individuals with dissociative seizures have epilepsy, which almost always preceded the onset of dissociative seizures. [68] Among people with a diagnosis of intractable epilepsy the incidence rate of dissociative seizures may be as high as 36% and leads to unnecessary treatment and the risk of ASM toxicity. [69]

People with epilepsy have a greater risk of death than the general population. Causes of death vary, including non-epilepsy-related conditions such as suicide, cancers, and cardiovascular disease, as well as epilepsy-related causes, such as status epilepticus, antiseizure medications (ASM)[2] effects, and accidents. [70] The three underlying causes of death in people with epilepsy reported by the National General Practice Study of Epilepsy were noncerebral neoplasm, cardiovascular, and cerebrovascular disease, accounting for 59% of deaths. Epilepsy-related causes, including sudden unexplained death in epilepsy (SUDEP) accounted for 3% of deaths. In 23% of individuals, the underlying cause of death was directly related to the epilepsy aetiology. [71]

SUDEP is an important risk in epilepsy that may affect about 1 in 1000 adults with the disease. [72]

One of the problems of studying SUDEP is the under reporting on death

[2] Previously referred to as anticonvulsant or antiepileptic drugs (AEDs) and this term may appear in quoted text throughout the thesis.

certificates. A review of sudden out of hospital deaths in Wake County (USA) found that 5.3% of the 399 death (18–64 age group) identified were due to SUDEP, but seizures or complications of seizures as the primary cause of death were only recorded on 1.5% of the certificates. [73] Similar under-reporting was identified in a review verified SUDEP cases in Sweden, where epilepsy was mentioned only on 63% of the death certificate. [74]

The incidence rates reported by this study was 1.20/1,000 person-years for definite/probable SUDEP, and was higher in men at 1.41 than in women at 0.96. All deaths in those aged under 16 years were in boys. The incidence rate increased 5-fold for women with psychiatric comorbidities compared to those without. [74]

Treatment

ASMs are intended to prevent seizure recurrence in people who have experienced unprovoked seizures and are the mainstay of treatment in epilepsy. There are numerous ASMs with a variety of mechanisms of action. Most of them operate by modulation of voltage-gated ion channels (sodium, potassium, calcium, chloride), by altering chemical transmission between neurons by affecting neurotransmitters (GABA, glutamate) in the synapses, by modulation of presynaptic neurotransmitter release, and by a combination of these mechanisms. [75]

In choosing the most appropriate ASM the seizure type has to be correctly diagnosed, but also patient's history, comorbidities, age, and sex have to be taken to consideration. Some of the most commonly used ASM, their target seizure types, and limitation are listed here. [76]

Sodium Valproate, with its mixed mechanism of action, is used for all seizure types, it is an enzyme inhibitor and is associated with teratogenicity and weight gain.

Carbamazepine (Na+ channel blocker), is used to treat focal seizures, is an enzyme inducer and is not useful for absence or myoclonic seizures, it may cause skin hypersensitivity.

Topiramate is used to treat focal and generalised seizures, it may produce cognitive side effects, kidney stones, speech problems, and weight loss.

Levetiracetam is used for focal and generalised seizures, and can be useful for absences and myoclonic seizures. It is associated with psychiatric side effects. Lamotrigine is used to treat focal and generalised seizures, it is associated with skin hypersensitivity.

Tolerability of treatment is essential to its efficiency and it is a significant issue with ASM with some 88% of individuals reporting at least one side effect. [77] These may include tiredness, memory problems, issues with concentration, hair loss, weight gain. Side effects may lead some people to discontinue their medication. [78] A third of individuals self-reported changing their ASM at least once due to their side effects in their treatment. [77]

The introduction of the new generation of ASMs has not significantly improved their overall tolerability. A longitudinal cohort study of 1795 individuals with newly diagnosed epilepsy in Scotland reported that 15.6% of ASMs were discontinued within 6 months due to intolerable adverse effects. Individuals who had to stop their treatment due to the adverse effects were at higher risk of being intolerant to their current ASM.The proportion of second-generation There was no significant difference between the older and the newer (second-generation) ASMs rates of the adverse effects. [79]

The Standard and New Antiepileptic Drugs (SANAD) study of lamotrigine, topiramate, valproate use in generalised and unclassifiable epilepsy reported at least one adverse effect in around 40% of individuals. They included tiredness, fatigue, weight gain/loss, personality change, worsening seizures, accidental injury, headache, memory loss, and depression, [80] Similar results were reported for lamotrigine, gabapentin, oxcarbazepine, and topiramate treatment for focal epilepsy with about 50% of patients reporting adverse effects. [81] A data linkage study using the SAIL databank reported an increased risk of major cardiovascular events in people with epilepsy [82]

Finally, ASMs may have intergenerational consequences. For example, in utero exposure to ASMs in combination, or sodium Valproate alone, was found to be associated with a significant decrease in educational attainment in national educational tests for 7-year-old children compared with both a matched control group and the all-Wales national average, supporting the notion of the cognitive and developmental effects of in utero exposure to Sodium Valproate as well as multiple

ASMs. [83]

Although significant progress has been made in the ASM developments with new targeted treatments being available the levels of drug-resistant epilepsy are still high [84]

In a 25 year follow up of 1098 individuals on ASM treatment 68% became seizure free (no seizures for more than a year). The pattern of seizure control among the individuals within the cohort was that 37% achieved early and sustained seizure freedom, 22% had delayed but sustained seizure freedom, 16% experienced fluctuation between periods of seizure freedom and relapse, and 25% were never seizure free. There was a higher probability of seizure freedom in patients receiving one compared to two drug regimens, and two compared to three regimens. [85] Other studies also suggest that a failure of two tolerated ASMs reduces the chances of success subsequent drug therapy. [86]

It has to be noted that there may be other determinants of treatment failure. The most common reasons reported in a study of over 300 newly diagnosed individuals with uncontrolled epilepsy was single ASM tried (56%), poor compliance (34%), adverse effects on a small dose (29%), iadequate dosing (28%) alcohol and/or recreational drug use (19%), psychiatric problems affecting documentation, attendance, etc. (18%) [87]

A multicentre study of over 800 individuals reported that forgetting to take ASM was associated with lack of seizure control (focal to bilateral tonic–clonic seizures). Dementia, younger age, use of multiple drugs, and living alone were identified as the risk factors. [88]

Impact

Epilepsy has an extensive and multifaceted impact on people with the disease and their families. People with epilepsy, apart from the physical aspects of having seizures, taking medication which may produce side effects, having injuries, and a higher rate of comorbid conditions, still experience a significant level of social stigma associated with the disease which has an effect on education and employment [89] Children with epilepsy and their parents experience social exclusion, activity restriction, and teasing/bullying. [90] The attitude of employers to individuals with epilepsy has not significantly changed with many (∼20%) still thinking that employing people with epilepsy may be '*major issue*' with concerns about safety and work-related accidents being expressed. [91] Epilepsy has been associated with psychological and emotional problems, social isolation, and problems concerning education, employment, family life, and leisure activities [92]

Epilepsy is a heterogeneous disease with complex aetiology including genetic and environmental influences that are still not clearly understood. [93] Numerous comorbidities may be a part of this aetiology, may affect treatment, and influence disease prognosis. Despite the advancements in treatment, some 30% of individuals still do not gain full seizure control, but the full picture of this failure is not clear. Could access to larger cohorts of individuals with epilepsy such as is offered by routinely collected health care information help to provide some better insight into this area?

1.3 Routinely collected healthcare data

Routinely collected healthcare data, that is the data primarily generated for administrative or surveillance purposes and not to answer specific research questions, has long been used for healthcare research. [94] Hospital administrative systems, disease registers, and primary healthcare records have been used in utilisation, outcomes, and epidemiological studies. [95–101] This research has been possible because of the structured nature of the data collected, dataset standards, and coding. Two classification systems within secondary care that enable patient based research are the International Classification of Diseases (ICD) with the 10th revision being used in the UK since 1995, applied for diagnosis coding of inpatient episodes. The second, the Classification of Interventions and Procedures (OPCS-4), is used for coding of inpatient interventions and surgical procedures. In primary care, READ codes, developed by Dr James Read and deployed across the UK since the mid-1990s, have been used in Electronic Health Record (EHR) to code patient's symptoms, presenting complaints, history, examinations, investigations, tests, preventative procedures, chronic disease monitoring, operations or other therapeutic procedures, and referrals. [102]. These are now being replaced by SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms).

As the use of these records in research increased so has the need to measure their reliability, with numerous studies aimed at assessing their accuracy and consistency, often identifying coding as an area of concern. A systematic review by Campbell et al., of studies comparing discharge statistics and the original medical records, reported median coding accuracy rates of 91% for diagnostic codes (range:74–97%) and 69.5% for operation or procedure codes in England or Wales, with 82% (range; 66–94%) and 98% respectively in Scotland. [103]. In a review from 2011, Burns et al. assessed a number of studies comparing routinely collected datasets to different reference data sources. [104] The studies using case or operative notes reviews for reference (n=25), reported a median diagnostic accuracy of 80.3% with significant levels of variation, and a median procedure accuracy of 84.2%. However, it was noted that since 2002, the diagnostic accuracy for primary diagnosis increased from 73% to 96%. The majority of inaccuracies as reported by the studies related to four digit coding, this suggests that three digit coding (which may be sufficient) would produce fewer errors. [104]

Jardan et al., reviewed a number of studies investigating the quality of GP morbidity registers by reference to paper notes, prescribing information or diagnostic tests and procedures. [105] Although it was difficult to generalize the findings, due to the variation in the methodology and quality of the studies reviewed, the common finding was that the quality of recording varied between comorbidities. Similar results were reported by Khan et al., in a systematic review of studies relating to diagnostic coding within the General Practice Research Database (GPRD). [106] Bearing in mind very different approaches taken by the studies (questionnaire, review of notes, verification against hospital letters) and different diagnoses being investigated, the reported positive predictive value (PPV) ranged from 21% to 100%. Human errors, computer systems that do not promote coding consistency, diagnostic uncertainty, and professional experience, all play a part in the quality and accuracy of clinical coding. [107–109]

A significant advantage of the EHRs is the ability of data linkage across different providers, disease registers, and administrate datasets. This allows for creation of large population based research cohorts ideal for rare disease and longitudinal studies. There are many successful population-based data linkage infrastructures and systems, differing in the range of data types (health and non-health) and the number of datasets available for linkage, facilitating a magnitude of studies. [110–116] In Wales, the Secure Anonymised Information Linkage (SAIL) databank was established in 2006 in recognition of the vast research potential of person-based data collected by health and other public services.

1.3.1 The SAIL Databank

The data linkage described in this project was performed within the SAIL Databank. Hosted by the Health Data Research(HDR) UK site at Swansea University, the SAIL Databank is a national data safe haven of anonymised health and administrative datasets about the Welsh population. The collection contains regularly updated health related datasets including primary and secondary care, births and deaths, health screening programmes (bowel, breast, and cervical cancers), health registers (congenital anomaly, cystic fibrosis), and social care services. There are also many administrative datasets, such as those relating to education and looked after children, and family justice datasets. [117]

The anonymity of the datasets is achieved by the separation of personal demographic information, such as name, date of birth, and NHS number from the clinical/administrative event details. There is a single common key (ID) that links the datasets. All data providers split their datasets along this division, with the identifiable data being sent to a trusted third party (TTP), this service is at present provided by the Digital Health and Care Wales, and the event data submitted to the SAIL Databank. The demographic data are anonymised, encrypted and checked, with each record being assigned an Anonymous Linking Field (ALF) or Residential Anonymous Linking Field (RALF) for places of residence. The datasets containing only the ALF, week of birth, gender code and area of residence, as Lower Super Output Area of approximately 1500 head of population, are then sent to the SAIL Databank. [117] The common key field is encrypted once the datasets are uploaded and is used in the linkage.

The datasets available within the Databank are stored as relational databases (IBM Db2), that can be interrogated using SQL queries, with additional applications available for further analysis, such as R, SPSS, Python. In addition to the datasets provided within the Databank the secure infrastructure offered allows researchers to upload their own data for linkage to the existing SAIL datasets, using the same split file method described above.

1.3.2 Epilepsy research using linked data

An essential starting point for epilepsy studies using EHRs and linked data is the correct identification of individuals with epilepsy. Validated case definition algorithms set a design pathway and give credibility to further studies. A number of approaches have been reported. An epilepsy diagnosis validation study using data linkage within the SAIL Databank and a cohort of individuals with confirmed epilepsy validated three algorithms for case ascertainment from GP records, concluding that the most reliable method was to select individuals who had a diagnosis of epilepsy and were prescribed ASM. [118] A systematic review of approaches to case ascertainment from administrative data assessed 30 studies, concluding that in order to achieve a high PPV an algorithm should combine disease and ASM prescription codes. [119]

There has been a number of data linkage studies investigating different aspects of epilepsy, some of which were performed using the SAIL Databank: mortality, [120–122] prevalence in known comorbidities, such as cerebral palsy(CP), [123] attention deficit hyperactivity disorder (ADHD), [124] schizophrenia, [125] multiple sclerosis (MS), [126] association of epilepsy and ASM with major cardiovascular events, [82] risk of dementia in people with epilepsy, [127] quality of care, [128] surveillance of pregnancy outcomes, [129] the association of epilepsy prevalence with deprivation,

[130] and the in utero exposure to ASM and educational attainment, [83] are some of the examples.

These studies demonstrate that routinely collected data can facilitate a wide ranging research but the challenge of obtaining sufficiently detailed diagnostic information that is needed for more in-depth studies still remains. This is because the datasets available are based on the structured and codified data produced by the primary and secondary care, without the free text element of these records being included. GP EHR contain free text notes, and clinic letters, discharge summaries, and test results provided by the secondary care are stored electronically and could be used given the appropriate tools.

1.4 NLP for Clinical Information extraction

Natural language processing (NLP) refers to technologies that enable computer programs to process and ‘understand’ human language. It is widely used in everyday applications (language translation, speech recognition, text summarising) and for the last few decades has been applied in extracting information from free text contained in clinical documents. [126] The conversion of a narrative describing patient’s history into a structured format enables the generation of clinical data not previously available, which may provide valuable support in clinical decision making and constitute a significant research resource, especially in combination with large population based datasets.

The approaches to building an NLP application are usually divided into rule-based and machine learning methods. [131] Rule based systems use standard NLP elements (part of speech taggers, regular expressions) combined with specialist lexicons, such as Unified Medical Language System (UMLS), to build patterns that are matched within text. They rely on domain knowledge and are usually developed for a specific disease or clinical concept (diagnosis, smoking status, pain status, cancer staging, and medication). Because of specialist clinical input required, rule-based systems are accurate but time intensive to develop. Machine learning approaches ‘recognise’ patterns that were previously marked as the information of interest (diagnosis, symptoms, test result). They have become more commonly deployed, and of those, Support vector machines (SVM), Conditional random field (CRF), Random forest, and logistic regression being some of the most popular. [132] These systems are efficient and effective but the lack of large annotated training sets limits their application. [133, 134] Hybrid systems which combine rule based and machine learning approaches have also emerged. [108]

A 2020 review of clinical concept extraction publications reported that 48% of systems used rule-based approach, and half of these for specific disease areas, machine learning and hybrid systems each represented 22% of studies, with 8% using deep learning methods. [135] An earlier review by Wang et al., suggested that rule-based models are still the most commonly used in the clinical extraction field, but the choice also depends on the task. [136]
One of key key advantages of the rule-based approach is the explicit and explainable decision process. Developing a rule-based model for a concept requires a logical understanding of its meaning in the medical context, its reporting style, and even the thinking behind the reporting style. For the same reason, it is important that NLP researchers and clinicians to work closely to refine the rules with explicit domain knowledge.

There are various tools and systems that can be used for clinical NLP tasks, some have been developed to address specific areas, others are more wide ranging in scope, and there are also general NLP systems that have been adapted to the clinical domain. Many of the developments relating to specific tasks were initiated by the challenges set out by the Informatics for Integrating Biology & the Bedside (i2b2). Funded by the National Centre for Biomedical Computing (NCBC) in the USA, the i2b2 made available deidentified annotated datasets for a number of clinical NLP challenges. [137] These tasks, which included deidentification, smoking, obesity, medication, temporal relations, and heart disease, set out clearly the type and format of the information to be extracted, but also used standardised validation methods.

An example of a single task development is a Mayo Clinic NLP system for smoking status identification. This is a classifier built on Unstructured Information Management Architecture (UIMA) and the SVM implementation for a specific task of identifying patient-level smoking status. [138] A medication information extraction system (MedEx) was developed specifically to extract prescription information from discharge summaries but it has performed well on outpatient clinic visit notes. [139]

The clinical Text Analysis Knowledge Extraction System (cTAKES) (now Apache cTAKES) was developed by Mayo Clinic using (UIMA) and OpenNLP natural language processing toolkit. [140].

The Clinical Language Annotation, Modeling, and Processing (CLAMP) toolkit was created as a flexible clinical NLP pipeline development application that can be used as a command line NLP system, to extract concepts built on default components, or with a graphical user interface (GUI) for building customized applications. [141]

Medical Language Extraction and Encoding System (MedLEE) designed for decision support and initially applied to radiology reports of the chest, was later extended to mammography reports and later to discharge summaries. [142] Stamford CoreNLP [143] and The General Architecture for Text Engineering (GATE) [144] are two example of general purpose systems that are used in the development of clinical NLP application. Health Information Text Extraction (HITEX), for instance, was built using the GATE framework to address the i2b2 tasks of extracting diagnoses, comorbidities, discharge medications, and smoking status from various types of medical records. [145]

The ExECT pipeline which is discussed in this thesis has been developed using GATE.

1.4.1 UMLS

Identification of medical concepts is an essential initial step in the development of NLP application. The Unified Medical Language System has been shown to be a comprehensive and valuable resource for clinical concept mapping and is widely used in clinical NLP. [146] The UMLS is maintained by the National Institute of Health (NIH) National Library of Medicine (NLM). It is essentially a collection of databases and tools to facilitate mapping between different biomedical and health related vocabularies. The vocabularies form a large database, the Metathesaurus, which links related terms from different sources under a single concept, with a Concept Unique Identifier (CUI). Preferred term (PREF) is designated to describe each concept 'based on an order of precedence of all the types of English strings in all the Metathesaurus source vocabularies'. [147] The UMLS is updated biannually in May (AA release) and in November (AB release), although not all vocabularies are updated at that rate.

1.4.2 Gold standard annotation

Setting out clear aims for an NLP algorithm by defining what information should be extracted and what format it should take is an important first stage of the development process. The creation of a gold standard annotation, apart from its set objective of creating a standard, allows for the aims of the algorithm to be clarified and if necessary redefined.

A gold standard annotation is a set of correct annotations developed by manually annotating text according to a set of guidelines by at least two trained annotators. [148] The annotations are compared, reviewed, and any disagreements are resolved, so that the final set represents the 'ground truth'that can be used in the development and validation of NLP applications.

The task of creating the gold standard for a specific project may be led by experts in the field, algorithm developers, and / or trained annotators. For example, in the work on common data elements in operative notes for knee arthroplasty, the gold standard was build by trained registry specialists collaborating with orthopaedic surgeons and data scientists. [149] The sets may be created by teams developing specific applications, or for the benefit of the wider clinical NLP community. The challenges set by the i2b2 are an example of gold standards being created for this purpose. [137,150,151] This initiative is now being continued by the Harvard Medical School Department of Biomedical Informatics who organise the National NLP

Clinical Challenges (n2c2).

The key element of a gold standard development are the annotation guidelines. They define the task of annotating to ensure consistency across the annotations. Depending on individual projects, the guidelines may be relatively simple, especially when the concepts being annotated are comprehensible and without additional features (attributes), [152] or very detailed, when the entities are complex and additional features or temporal relations are involved. [153]
Many studies reporting on the development of a gold standard do not provide very detailed information about the annotation schemas used. One of the exceptions is the annotation methodology described by Roberts et al., which reviews in depth the development of annotation guidelines as part of the Clinical E-Science Framework (CLEF) project. [154] This work involved multiple clinician annotators and computational linguists, went through many development iterations, and the resulting website contained easy to navigate sections with definitions, examples, and specific guidelines for each entity and relations.
The number of annotators required varies depending on the corpus size, the complexity of the task, and the systems used. For the extraction of principal diagnosis, comorbidity, and smoking status for the HITEX tool project, a single asthma expert reviewed 150 discharge summaries, annotating the content for five concepts with 'yes', 'no', and 'insufficient data' with the decisions being confirmed by four other physicians. [155] In the Development of a validation corpus to support the automatic extraction of drug-related adverse effects from medical case reports, three annotators were involved in the annotation of entities and relations in 2000 documents each. [156]

The annotation workflow usually involves a number of trial tests, to assess the difficulty of the task and to clarify or adjust the guidelines, and the measurement of the Inter Annotator Agreement (IAA). The latter is used to assess the level of agreement between the annotators and may reflect the difficulty of the task. The IAA can be measured by Cohen's kappa, for classifier applications with known number of entities, or F1 score, for named entity recognition.

Cohen's kappa [157] measures the agreement between two annotators taking to consideration the portion of agreement that can be achieved purely by chance. In other words, it compares the probability of the two annotators agreeing by chance to the observed agreement, which can be expressed mathematically as:

$$k = \frac{Po - Pe}{1 - Pe} \tag{1.1}$$

where *k* is the kappa value, *Po* is the proportion of the observed agreement and *Pe* is the proportion of expected agreement, hence *Po-Pe* represents the proportion of the cases in which beyond-chance agreement occurred [158]

F1 score is based of on a contingency table that identifies true positives (TP), false positives (FP), and false negatives (FN) when comparing annotations produced by the annotators, using one set of annotations as the ground truth.

$$Precision = \frac{TP}{TP - FP} \tag{1.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{1.3}$$

where TP = True positive is a number of correctly extracted annotations, FP = False positive is a number of incorrectly extracted annotations, and FN = False negative is a number of missed annotations.

The harmonic mean of precision and recall is then calculated with the general formula as

$$F = \frac{(1 + \beta^2) \times recall \times precision}{(\beta^2 \times precision) + recall} \tag{1.4}$$

Where β is a factor by which precision or recall can be weighted if either of them is considered to be more important. Otherwise it is set as 1 as in

$$F1 = 2 \times \frac{recall \times precision}{precision + recall} \tag{1.5}$$

A low score, for example less than 0.6 (F1, lower for Cohen's kappa), indicates poor agreement, which may be due to the task being difficult or that the annotators have not been properly instructed. [159]

During the creation of the corpus for the i2b2 smoking status challenge, described by Uzuner et al., [137] the smoking status of patients recorded in their discharge summaries was annotated by two pulmonologists for specific categories of 'past smoker', 'current smoker', 'smoker'(past or current), 'non-smoker'(never smoked), and 'unknown'. Although the coverall agreement measured by Cohen's kappa was 0.84, for specific categories it ranged from 0.4 to 0.98, suggesting a significant disagreement on some of them. Additional pulmonologists were invited to adjudicate, but as for some cases no agreement could be reached these records were removed from the set. For this reason the IAA may be measured a number of times before the final annotation of the validation test is made and consensus sessions are essential to arrive at the ground truth.

1.5 NLP for epilepsy

One of the most striking issues about epilepsy in the fact that with the advances in drug development around one third of individuals do not become seizure free, despite treatment. This may be partly related to incorrect diagnosis, interaction of comorbidities, and/or aetiology, or other factors which are not clearly understood. It is hoped that with the availability of more detailed diagnostic information, seizure description, frequency of events, detailed information of ASM dosage, comorbidities, and full history for large populations of individuals with epilepsy some patterns will emerge that will help to unravel the complex interactions of all these factors.

Using routinely available information such as GP, hospital and administrative data significantly aids epilepsy research. [25, 82, 83, 160] It does not, however, offer the level of detail that is needed to enrich phenotypic information or gain in-depth knowledge of individuals' health. For example, age of onset (significant for correct diagnosis and treatment prognosis), ASM dosage, seizure type and frequency, detailed comorbidities, are not available from routinely collected GP or hospital data that are available from electronic systems. Free text information, such as clinic letters, discharge summaries, or test results is, however, stores electronically and could be accessed.

Epilepsy is particularly well suited to NLP based information extraction systems for its reliance on a descriptive diagnostic process, based on history provided by patients and/or their relatives, friends or carers, which may be supported by EEG and/or neuroimaging. Patient accounts together with a commentary by a clinician, and other details relating to symptoms and treatment are contained in a narrative of clinic letters. These, along with the investigation reports contain a remarkable amount of detail that could greatly enrich routinely collected data.

An important aspect of developing an NLP application for epilepsy is the availability of ontology that adequately reflects the recent developments in classification of the diagnostic terms but also captures the general language used in clinical text. This may present a considerable difficulty as any system would have to capture the new and the old lexicon. For example, it may be challenging for a system to identify focal or generalised seizure type if some of the terms are used without clarification, based on a clear diagnosis, i.e., myoclonic seizure instead of focal myoclonic seizure when it is given in the context of focal epilepsy.

1.6 Review of existing systems

There have been some developments of systems designed specifically to capture clinical information relating to epilepsy. Epilepsy data extraction and annotation

(EpiDEA) described by Cui et al., is a rule-based system, which extracts information from epilepsy monitoring unit discharge summaries. EpiDEA achieved an overall precision, recall and F1 score of 0.94, 0.84 and 0.89, respectively, when extracting EEG pattern, past medications and current medication from 104 discharge summaries from Cleveland, Ohio, USA. [161]
Phenotype extraction in epilepsy (PEEP) is another rule-based algorithm developed to extract epileptogenic zone, seizure semiology, lateralising signs, interictal and ictal EEG pattern. In the validation process without accounting for the location information it achieved micro-average precision, recall and F1 score of 0.93, 0.93 and 0.92, respectively. The results were slightly lower with the exact match of the epiletogenic zone location. [132]

A machine-based learning NLP pipeline to identify West syndrome from discharge summaries and EEG reports was developed in Phoenix (USA). Two different feature generation methods were used, term frequency-inverse document frequency (TF-IDF) vectors and a topic distribution based. The best results were produced by Support Vector Machine classifier with TF-TDF combination and achieved precision, recall and F1 score of 0.77, 0.67 and 0.71, respectively. [162]

There have been some developments in using NLP for capturing seizure frequency information. An NLP algorithm to extract frequency expressed in quantitative values (daily, 3 per week) has been described by Decker et al. It achieved precision of 0.95, recall of 0.70, and F1 score of 0.81. for internally annotated test set. For externally annotated test set, the results were lower, at 0.73, 0.22, and 0.40 for the three measures. [163]
In another development, three models were used to extract seizure frequency (BERT, Bio_ClinicalBERT, and RoBERTa), expressed as a classifier 'has the patient had recent seizure', as a quantifiable variable 'How often patient has seizures' and, a temporal measure, 'when was the most recent seizure '. For the classification tasks the first two models achieved 80% accuracy, and all three produced F1 score of 0.86 and 0.85, respectively for the extraction of frequency and date of last seizure. [164]

There have not been developments that produced structured datasets of a whole range of epilepsy information extracted from clinical text, apart from the ExECT (extraction of epilepsy clinical text) pipeline created by our team. [165] ExECT v1 was built using the GATE framework, with its biomedical named entity linking pipeline (Bio-YODIE), and a customised version of the South London and Maudsley (SLaM) GATE application to extract prescription information. [166] The ExECT pipeline extracted nine epilepsy categories: epilepsy diagnosis and type, focal seizures, generalised seizures, seizure frequency, medication, and investigation results (CT, MRI and EEG). In a 200-letter test it produced the overall precision, recall, and F1 score of 91.4%, 81.4% and 86.1%, respectively. This thesis describes the redevelopment of the pipeline, and some comparisons will be made between the two versions in the following chapters.

1.7 The genetics of outcomes in epilepsy

Many factors may influence an individual's response to ASM therapy but it is highly likely that genotype has a strong effect. SUDEP, the most severe epilepsy outcome, has been shown to be associated with an increased polygenic burden and a greater presence of potentially deleterious variants, but no single gene has been identified as common to the SUDEP cases. [167]

There is some evidence of association between genetic markers and response to ASMs (specific drugs or as general pharmacoresistance) from investigations into all or particular epilepsy types. Pharmacogenomic studies have identified genes associated with poor response to ASM, such as mutations in encoding CYP enzymes, transporter genes, and genes associated with seizures that are also linked to pharmacoresistance, such as *SCN1A*. [168]
Common polymorphisms in T-type calcium channels*CACNA1G, CACNA1H, CACNA1I* and 1 transporter variant *ABCB1* were found to be associated with seizure outcomes in a group of 446 children newly diagnosed with CAE treated with ethosuximide, lamotrigine, and Valproate. 2 polymorphisms (in *CACNA1H, CACNA1I*) were more common amongst non-seizure free participants on ethosuximide, 1 polymorphism (in*ABCB1*) was seen more commonly in non-seizure free individuals

on lamotrigine, with 2 in*CACNA1H* being more common in the seizure free individuals. There were no associations between common polymorphisms and seizure status in participants on Valproate. [169]

A study of patients with GGE used genome-wide analysis to investigate the influence of common and rare genetic variants on their response to lacosamide. It failed to identify any variants that were specific to people experiencing 75% seizure reduction or seizure freedom, 73 participants, or those who experienced less than 25% reduction in seizures, 495 individuals. [170] Another genome-wide association study investigating ASM response in GGE assessed 3.3 million SNPs in 893 individuals treated with lamotrigine, levetiracetam, and Valproate. There were no significant genome-wide markers identified in responders (individuals achieving seizure freedom for at least one year) and non-responders (those with recurring seizures at 50% or above of pretreatment frequency level), although 29 loci were possibly associated with ASM response. [171]

The role of rare variants in resistance to specific ASMs was assessed using genetic burden analysis for 1622 whole exome sequenced individuals treated with lamotrigine, levetiracetam, or valproic acid. Rare missense and truncating variants in genes involved in valproic acid pharmacokinetics were enriched in individuals who were resistant to valproic acid as compared to those who were not. There was also some enrichment in truncating variants in synaptic vesicle glycoprotein (SV2) family genes in individuals who were resistant to levetiracetam. There was no significant enrichment shown in gene-based analysis. [172]

A much larger study of 3,649 individuals with focal (2,768) and generalised epilepsy (887) looked at the association of common genetic variants with response to individual or groups of related ASMs. No significant genome-wide association was identified, although 30 loci were suggestive of potential involvement in ASM response. [173]

Deleterious variants

There are a number of systems to predict the potential pathogenicity of genetic variants. For example, PolyPhen scores predict the effect of an amino acid substitution on the structure of protein, providing a score and a result as ‘probably damaging, possibly damaging, benign, unknown '. SIFT predicts whether an amino acid substitution is likely to affect protein function and is based on sequence homology and the physico-chemical similarity between the alternate amino acids. The score (annotations) produced is either 'tolerated'or 'deleterious'. [174]

Combined Annotation-Dependent Depletion (CADD) is a framework integrating multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. It provides the scores of deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome. [175] Raw scores can indicate how likely it is that the variant has derived from the proxy-neutral (negative values) or proxy-deleterious (positive values) class. PHRED scores (scaled) are derived from the relative ranking of model scores across all potential SNVs. [176]

1.7.1 Objectives

The aims of this thesis were to redevelop the ExECT pipeline to produce detailed datasets of epilepsy information that could be linked to routinely collected data and used in epilepsy research.

The main elements of this project include:

- The creation of a gold standard dataset for epilepsy clinic letters annotated by trained annotators according to collectively developed and agreed guidelines.

- Create and validate a pipeline for extracting identifiable data from clinic letters that can be used to create a file of personal demographic data for encryption and linkage.

- Redevelop the ExECT pipeline by constructing up-to-date gazetteers, and rules to capture an expanded range of variables in a standardised format.

- Validate the output produced against the gold standard developed in the first stage of the project and compare with the original results achieved by ExECT.

- Apply ExECT and IDEx on a large set of clinic letters for a cohort of patients from the Epi25 study and produce a structured output.
- Validate a selection of the variables extracted against the Swansea Neurology Biobank dataset.
- Process and link a structured output from seizure frequency and prescriptions and analyse for a selection of individuals and for the entire cohort.
- Annotate genetic data from the Epi25 cohort within the SAIL databank and link with the routinely collected GP and hospital datasets (Genetic linkage study).
- Upload and link seizure frequency dataset created by ExECT to the Genetic linkage study.

Chapter 2

MATERIALS AND METHODS

This chapter describes materials and methods used to create and validate the NLP pipelines for annotating clinic letters, to process the output created, and to use it in a data linkage study.

2.1 Document sourcing and preparation

Throughout this project clinic letters and a small number of other clinical documents, such as test results, from hospital adult and paediatric neurology services were used. All of them relate to the Swansea Neurology Biobank (SNB) donors who have consented for their clinical data to be used for epilepsy research (see section 2.1.3). Most of the documents were sourced from the Swansea Bay University Health Board (SBUHB) with a small number obtained from other Welsh health boards. Depending on the project stage they all required different levels and types of preprocessing, including pseudospeciation and de-identification, reformatting, or splitting.

2.1.1 Development and test sets

A development set is a collection of documents used to develop NLP algorithms. The documents should be representative of the language of the domain that is being studied and balanced in terms of different types of text. [148] For epilepsy clinic letters this means a selection of letters from several clinicians and ideally from different hospitals and health boards. They should relate to individuals with different types of epilepsy and at various stages of diagnosis and treatment i.e. first seizure clinic, routine follow-up appointments, reviews of more complex epilepsy, referrals to other services, or letters containing investigation results, as all of these different types of letters use different styles and content.

Extraction of Epilepsy Clinical Text (ExECT) v2 was developed and tested on a 200-letter corpus used for validation of ExECT v1 (section 1.6) These documents were manually psuedonymised with all personal and demographic information such as name, date of birth, address, school or work place, and employment being substituted with fictitious terms. Hospital, Clinician, GP, and clinic information were also changed. Any letters containing other information that could potentially be identifiable or was considered sensitive were removed from the set.
The same process was used in the ExECT v1 40-letter development set which was reused in the construction of IDEx algorithms.

2.1.2 Gold standard annotation development and validation sets

Letters which were used for Markup (section 2.3.2) annotation tests and gold standard set were sourced from the SNB records. These were a subject to the same pseudonymisation and de-identification as the development set.
For Markup annotation all documents had to be converted to plain text and processed to remove formatting i.e. tabs or white spaces before creation of .ann files. Figure 2.1 A gives the examples of scripts used.

```
A

# 1 Java tikka application converting word documents to plain text files
java -jar tika-app-1.20.jar -t -i InputFolderName -o OutputFolderName
# 2 Finding and replacing TABS with Spaces in all files in a given folder / directory
     saving the original file with an .orig extension
find . -type f -exec sed -i.orig 's/\t/    /g' {} +
# 3  Deleting all blank lines before first word
find . -name "*.txt" -type f -exec sed -i -n '/[^\t]/,$p' {} +;
# 4 Creating .ann files from text files  for BRAT and Markup
for f in *.txt; do
touch "$f" "$(basename "$f" .txt).ann"
done

=============================================================================================

B
# 5 Splitting multiple letters in a single word document into one letter per document using
       specific phrases e.g. Neurology Department ( Heading) or Your sincerely
$ csplit --prefix NewFile_ OriginalFile.txt '/Phrase on which to split/' {*}
# Example splitting Bio_XXX.txt where a heading: NEUROLOGY DEPARTMENT appears with new
     files named Bio_XXX they will have numbers 01, 02, 03 ...depending on the number of
     headings, this is case sensitive so it will not split on /Neurology department/
$ csplit --prefix BioXXX_ Bio_XXX.txt '/NEUROLOGY DEPARTMENT/' {*}

# Here.txt suffix is added at the end of each file and the split in based on a string "XXX"
$ csplit --prefix=BioXXX_ --suffix-format="%d.txt" XXX_z.txt '/XXXX/' "{*}"
```

Figure 2.1: *Document conversion, preparation, and cleaning; A–preparation of single documents to be used in Markup: conversion and cleaning. B–examples of split points to separate individual clinic letters when they are grouped by patient in single files.*

Epi25 cohort document preparation

Clinic letters used for extracting information for the longitudinal linkage example (Chapter 6) and for the Genetic data linkage project (Chapter 7) were derived from the SNB records of consented SNB patients, held as PDF or Microsoft word format files, with some containing multiple letters which had to be split into single documents using bash script. PDF files were converted to plain text documents using Apache Tika, Fig.2.1 B shows the scripts used.

All documents were saved with an identifier based on the biobank number and a consecutive letter number for each person's set. The final cohort of 771 documents (from 1 to 48 per individual) consisted of clinic letters, letters to patients, letters to GP providing investigation results, and a small number of file notes containing EEG results.

2.1.3 Swansea Neurology Biobank database

The SNB database contains personal, clinical, and processing information relating to DNA samples collected and stored in the biobank. The Biobank has been approved by the Welsh Research Ethics Committee (REC 17/WA/0290). Donors are recruited from the NHS neurology clinics and provide written consent for their information to be collected and stored. This includes information gained from interviews and extracted from clinical notes. All individuals who participated in the Genetic data linkage study (Chapters 6 and 7) had also given consent to share their genetic and clinical data anonymously with the Secure Anonymised Information Linkage (SAIL) databank. Copy of the consent form used is given in Appendix A.1

SNB database

The SNB database is built in Microsoft Access 2007 and consists of two separate table collections, clinical records, held in NewBiobank.accdb, and donors' details in BiobankDonors.accdb. Biobank number is a common key field facilitating linkage. The database was used to extract information allowing for the second validation of IDEx and the evaluation of ExECT output from the Epi25 cohort set (Chapters 5 and 6).

The clinical database, apart from the sample record, contains disease specific, symptomatic, diagnostic, and treatment information. Epilepsy section content is shown in Appendix A2.1.3. The donors table contains personal demographic information and identifiable data that can be used for record linkage. Two separate datasets were produced using an Access query. The donors table provided a set of personal details to validate the IDEx pipeline (Chapter 5) and to create File 1 for the Genetic data linkage study (Chapter 7). The clinical dataset was used to validate diagnosis, investigations, and febrile seizures and to compare ExECT output for Onset and selected items from patient history (Chapter 6).

The SNB clinical dataset extracted The clinical dataset was extracted into a single csv file which was than split to Diagnosis and Seizures, Comorbidities,

ID	PREF	CUI
1	Bilateral convulsive(Secondary generalised)	C0877017
1	Focal dyscognitive(Comple xpartial) Focal	C0149958
1	Focal	C0014547
2	Bilateral convulsive(Secondary generalised)	C0877017
2	Absence-typical	C4316903
2	Tonic-clonic seizures	C0494475
2	Generalised	C0014548
3	Bilateral convulsive(Secondary generalised)	C0877017
3	Focal dyscognitive(Complex partial)	C0149958
4	Temporal	C0014556
4	Tonic-clonic seizures	C0494475
4	Focal	C0014547
4	Mesial Temporal Lobe Epilepsy with Hippocampal Sclerosis	C4749367

Table 2.1: *An example of the diagnosis and seizure information extracted from the SNB for a small number of individuals. The extract has been converted into a data table with the IDs based on the biobank numbers that could be linked to the ExECT Epi25 cohort output (substituted here with surrogates) and the UMLS CUIs assigned to the terms used in the SNB.*

and Investigations. As single columns contained many distinct diagnostic terms, they were split using the ‘Text to Column’ operation into separate columns in Microsoft Excel. As the validation of the ExECT output for the Epi25 cohort was performed using the Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) coded terms, phrases in the SNB output had to be appropriately coded. This was done manually, as the diagnostic terminology used in the SNB is different to the terms in the UMLS. This process achieved the output structure for Diagnosis and seizures shown in table 2.1.3 with substituted identifiers (SYSTEM_IDs). A similar format was produced for Comorbidities and Onset, which is illustrated with a short extract in Table 2.1.3. Investigation results were not assigned a CUI as the outcomes were classed as ‘Normal’ or ‘Abnormal’ and they were validated using these phrases.

IDEx and ExECT output validation was carried out in R Studio Version 1.4.1717 [177] using sqldf and dplyr packages.

ID	AGE ON-SET YEAR	AGE ON-SET MONTH	PREF	CUI
1	20		Bilateral convulsive(Secondary generalised)	C0877017
1	20		Focal dyscognitive(Complex partial) Focal	C0149958
1	20		Focal	C0014547
2	76		Bilateral convulsive(Secondary generalised)	C0877017
2	76		Focal dyscognitive (Complex partial)	C0149958
3		18	Tonic-clonic seizures	C0014548
3		9	Photosensitive response seizures	C0347873

Table 2.2: *An example of the seizure onset information derived from the SNB for a small group of individuals. The extract has been converted into a data table with the IDs based on the biobank numbers that could be linked to the ExECT Epi25 cohort output (substituted here with surrogates). The UMLS CUIs have been asigned to the terms used in the SNB*

2.2 Gold Standard development

A gold standard set is a corpus of documents annotated by the domain experts, setting out the entities of interest, features and the relationships to be extracted by an NLP application [178]. A subset of the annotated corpus can be used to guide the development and testing of the NLP algorithm, with a previously unseen larger set to be used for the application evaluation. [179] Inter-annotator agreement produced in the process of creating the gold standard provides a benchmark against which the performance of the application can be measured.

2.2.1 Annotation guidelines for Epilepsy

The development of annotation guidelines should follow an iterative process, where the scheme is developed, tested for reliability, the results are analysed to revise the scheme, and the process is repeated until the desired level of reliability has been

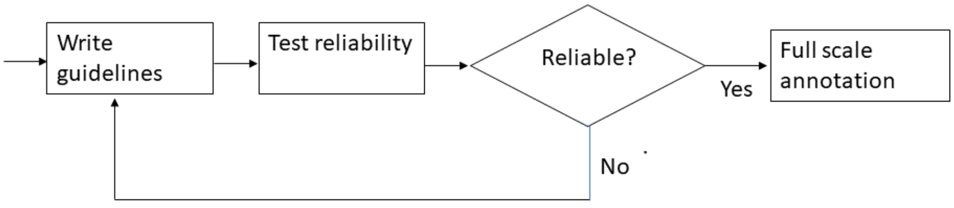


Figure 2.2: *Annotation Guidelines Development, reproduced from [180]*

Annotation tasks were performed using annotation software, Markup (section 2.3.2) with the guidelines setting out the entities and features to be annotated, explaining the process with numerous examples, and providing a list of terms and definitions. Suggestions from the annotators during annotation tasks and the IAA results led to multiple revisions (Chapter 3).

2.2.2 Inter-annotator agreement

Inter-annotator agreement (IAA) is used to measure the agreement between annotators and to assess the level of difficulty and complexity of the annotation task. [159] It identifies problems and can help in creating better guidelines that insure the uniformity and consistency of the annotations. [181] It can also be used as the benchmark for the performance evaluation of an NLP application. [156] Inter-annotator agreement was calculated at different stages of the manual annotation process, using different approaches, depending on its purpose.

Kappa statistic

There are a number of methods to measures IAA. The most commonly used is Cohen's kappa [157] which has been described in the introductory chapter (Section 1.4.2). It measures the agreement between two annotators taking into consideration the portion of agreement that can be achieved purely by chance.

An extension of Kappa developed by Fleiss [182] which allows for the calculation of IAA for more than two raters was used in this project and is described in more detail in Chapter 3, section 3.2.1. Kappa can take a value of -1 to +1, with negative values indicating that the agreement is less than expected.

The choice of scale in the interpretation of Kappa coefficient is arbitrary but it roughly follows the following ranges [183, 184]:

Kappa	**Agreement**
0.80-0.61	Almost perfect
0.60-0.41	Moderate
0.40-0.21	Fair
0.20-0.01	Slight agreement
≤ 0	Poor

Fleiss' kappa was used in the annotation tests to assess the change in the level of agreement and before the final annotation task setting the benchmark results for the validation process. Four annotators took part in the tests and the change in agreement was measured.

Precision, recall, and F-measure

The kappa statistic is suitable for classification tasks, where the annotators are charged with assigning defined categories (ratings) to a known set of entities, and it measures the agreement or disagreement between the annotators on each category. [185] It is not suitable for information extraction tasks when the number of entities to be annotated is not known. [186] Instead, precision, recall, and F-measure (F1 score) are calculated. [179, 187] The method follows the approach used for validation of NLP algorithms and it is described below, section 2.5, in reference to the evaluation of ExECT v2. It has also been described in the introductory chapter, in section 1.4.2.

The main difference is that the scores are calculated pair-wise, with each annotator's set being treated, in turn, as a key set. For the benchmarking IAA, the first stage of this process was performed within the GATE developer. Three annotation sets were uploaded using Groovy script, and precision, recall, and F1 score were produced for each pair. The results were then exported and the final calculation of the average scores was performed in Microsoft Excel for each annotated entity. The process of pair-wise IAA for the 20-letter annotation test is fully described in section 3.2.2. It results produced were used to provide a benchmark measure for the ExECT v2 validation (Chapter **??**, section 4.2).

2.3 Annotation software

At the onset of the project BRAT annotation software was chosen to create annotations on the validation sets. As the work progresses, an in-house system was being developed, and the creation of a gold standard annotation set for ExECT v2 became a testing ground for MARKUP.

2.3.1 BRAT

BRAT (brat rapid annotation tool) is an open source web-based annotation application that runs in a UNIX-like environment. [188]. The configuration file, annotation.conf, sets out entities, attributes, and relations for the phrases being annotated. Documents that are annotated have to be in text files (.tex). Annotations are saved in a standoff format (separately from the corresponding text file) in files with the .ann suffix. Annotations have a unique ID, specified by the following convention,

- T – entities
- E – events
- A – attributes
- R – relations

(N for normalisation was added in later versions) which are numbered, and are followed by numeric strings that define the annotation span (start and end).

BRAT was run in LINUX (UBUNTU) for annotation of the first IDEx validation letter set. A simple configuration set out the attributes for the four entities extracted to be based on the string being annotated, as shown below:

```
[entities]
Date_of_Birth
NHS_number
```

```
Pt_PostCode
Hosp_number

[events]
[attributes]
Date Arg:Date_of_Birth, Value:<ENTITY>
NHS_Number Arg:NHS_number, Value:<ENTITY>
PostCode Arg:Pt_PostCode, Value:<ENTITY>
Hosp_Number Arg:Hosp_number, Value:<ENTITY>

[relations]
```

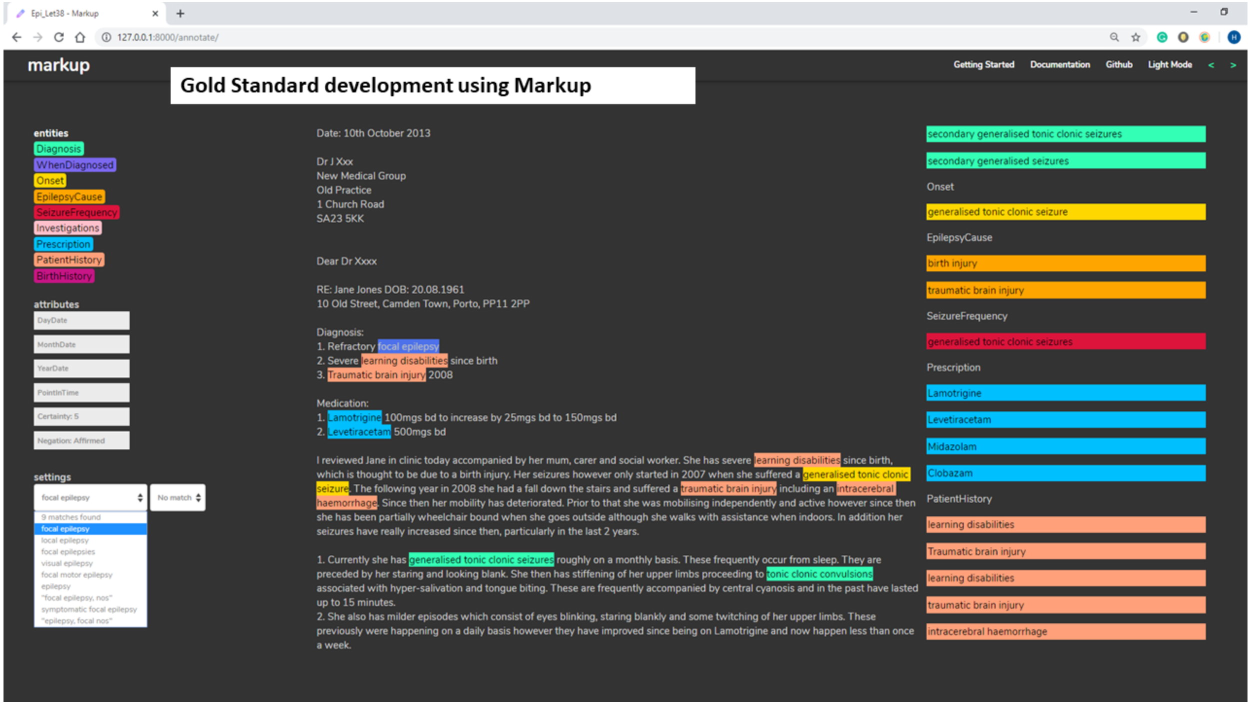
This produced the following .ann output, which could be uploaded into GATE to be used as a key set for IDEx output validation:

```
T1  Hosp_number 53 72 Hospital No. 112233
A1  Hosp_Number T1
T2  Date_of_Birth 212 227 dob: 11.11.1987
A2  Date T2
T3  NHS_number 228 248  NHS No. 604 604 6044
A3  NHS_Number T3
T4  Pt_PostCode 271 279 BA32 2WA
A4  PostCode T4
```

2.3.2 Markup

Markup [189] is an open source web-based annotation tool. It was constructed during the redevelopment of ExECT and was used in the annotation tests and the creation of the gold standard set. It is now available at https://www.get-markup.com/, but for this project a local installation was used. A configuration file defining all entities and attributes to be annotated was placed in the same folder as the plaintext files. For each text file an empty annotation file (.ann) was created in brat standoff format, using the following bash script:

Markup has an option for users to upload custom ontology which is used for automated term mapping. For ExECT v2 annotations a custom UMLS list was constructed by merging Epilepsy, Seizures, Comorbidities, Drugs, and Investigations gazetteers, stripping any classification groupings, and keeping only the UMLS term (PREF) and CUI.



Gold Standard development using Markup

Date: 10th October 2013

Dr J Xxx
New Medical Group
Old Practice
1 Church Road
SA23 5KK

Dear Dr Xxxx

RE: Jane Jones DOB: 20.08.1961
10 Old Street, Camden Town, Porto, PP11 2PP

Diagnosis:
1. Refractory focal epilepsy
2. Severe learning disabilities since birth
3. Traumatic brain injury 2008

Medication:
1. Lamotrigine 100mgs bd to increase by 25mgs bd to 150mgs bd
2. Levetiracetam 500mgs bd

I reviewed Jane in clinic today accompanied by her mum, carer and social worker. She has severe learning disabilities since birth, which is thought to be due to a birth injury. Her seizures however only started in 2007 when she suffered a generalised tonic clonic seizure. The following year in 2008 she had a fall down the stairs and suffered a traumatic brain injury including an intracerebral haemorrhage. Since then her mobility has deteriorated. Prior to that she was mobilising independently and active however since then she has been partially wheelchair bound when she goes outside although she walks with assistance when indoors. In addition her seizures have really increased since then, particularly in the last 2 years.

1. Currently she has generalised tonic clonic seizures roughly on a monthly basis. These frequently occur from sleep. They are preceded by her staring and looking blank. She then has stiffening of her upper limbs proceeding to tonic clonic convulsions associated with hyper-salivation and tongue biting. These are frequently accompanied by central cyanosis and in the past have lasted up to 15 minutes.
2. She also has milder episodes which consist of eyes blinking, staring blankly and some twitching of her upper limbs. These previously were happening on a daily basis however they have improved since being on Lamotrigine and now happen less than once a week.

secondary generalised tonic clonic seizures
secondary generalised seizures
Onset
generalised tonic clonic seizure
EpilepsyCause
birth injury
traumatic brain injury
SeizureFrequency
generalised tonic clonic seizures
Prescription
Lamotrigine
Levetiracetam
Midazolam
Clobazam
PatientHistory
learning disabilities
Traumatic brain injury
learning disabilities
traumatic brain injury
intracerebral haemorrhage

Figure 2.3: *Markup annotation tool screen, with ExEXC v2 configuration, showing annotations for diagnosis, comorbidities, and medication and an open UMLS drop down box.*

During annotation, Markup automatically maps any UMLS terms highlighted to the uploaded ontology and provides suggestions in a drop-down list. Terms can also be searched for should no suggestion appear. The left hand side of the screen displays entities and attributes (Fig.2.3) , with a drop down list of options (not visible) for each, on the far right side annotations that have been created are displayed, with a different colour for each entity. Annotations are saved to the already created .ann files showing entities (T) with the start and end strings, annotated phrase, and attributes (A) with the value selected from the drop-down list.

For example, in Fig.2.4, T22 is the entity of Onset, annotated in the phrase *'her epilepsy started when she was a young child'*, A54 to A60 are the attributes assigned to T22 by the annotator, with the age *'young child'* being given a numerical range of 1 for lower age (AgeLower) to 6 for upper age (AgeUpper), and epilepsy linked to the UMLS PREF (CUI Phrase in Markup) and CUI, in A59 and A60 respectively.

```
T105  PatientHistory 952 960  seizures
A245  Certainty T105 5
A246  Negation T105 Affirmed
A247  CUIPhrase T105 seizures
A248  CUI T105 C0036572
T20 Diagnosis 860 868 epilepsy
A48 DiagCategory T20 Epilepsy
A49 Certainty T20 5
A50 Negation T20 Affirmed
A51 CUIPhrase T20 epilepsy
A52 CUI T20 C0014544
T22 Onset 859 868 _epilepsy
A54 AgeUnit T22 Year
A55 AgeLower T22 1
A56 AgeUpper T22 6
A57 Certainty T22 5
A58 Negation T22 Affirmed
A59 CUIPhrase T22 epilepsy
A60 CUI T22 C0014544
```

Figure 2.4: *An example of the Markup annotation output for a clinic letter containing annotations with attributes for Patient History, Diagnosis, and Onset*

2.4 Algorithm development

General architecture for text engineering (GATE) https://gate.ac.uk was used to develop and validate the ExECT v2 and IDEx NLP pipelines, and to produce datasets based on the annotations. GATE is an open source software toolkit providing an infrastructure to built text processing and analytics applications. [190] Both pipelines were created in GATE Developer, the GATE graphical user interface (GUI), with the core language and processing resources (PR) such as tokeniser, sentence splitter, part of speech tagger (PST), semantic tagger, ConText algorithms, and user defined components.

2.4.1 Tokeniser

Tokeniser annotates every token, be it a space, word, number, or punctuation, with a set of attributes (excluding space tokens that are given length only):

- category – assigns part of speech tag such as JJ for adjectives, NP for noun singular, PP for personal pronoun
- kind – word, number, punctuation, or space
- length – string length
- orth(orthography) – such as upperInitial, allCaps, lowerCase
- string – string itself

Tokens are the smallest building block used in rule building and are useful for capturing specific patterns such as postcodes. They were used in IDEx for rules annotating Hospital and NHS numbers.

Sentence splitter divides text into individual sentences and is helpful in restricting rules to a single sentence.

2.4.2 Named entities – Gazetteers

Named entities are identified in text and assigned features when they are matched against terms held in lists, stored in collections referred to as gazetteers. Lists can be very short or very long, and can assign different features to the terms contained. For example a list of terms used as triggers to identify individual's date of birth in IDEx has 8 simple entries of:

D.O.B
d.o.b.
date of birth
Date Of Birth
Date of Birth
dob
DOB

DoB

On the other hand, epilepsy list, which holds diagnostic terms for epilepsies and seizures, has 946 entries, as it attempts to capture every variation of specific epilepsy type or seizure that may appear in clinic letters. When features are added to the items in a list they are separated by a tab. A list collection is stored in .def file with a one-line record for each list defining the annotation type it produces (Lookup) as illustrated below by IDEx.def, for which all gazetteer annotations are stored in Lookup2:

```
nhs.lst:person:health_term:NHS:Lookup2
hosp.lst:person:health_term::Lookup2
birth.lst:person:date_term::Lookup2
daydate.lst:Numerals:Ordinals::Lookup2
months.lst:time:date::Lookup2
letter.lst:reference:ref_term::Lookup2
consult.lst:reference:health_term::Lookup2
patient.lst:person:ref_term:gender:Lookup2
person_female.lst:person:female:gender:Lookup2
person_male.lst:person:male:gender:Lookup2
title_male.lst:title:male:gender:Lookup2
title_female.lst:title:female:gender:Lookup2
ClinicDate.lst:ClinicDate:Date::Lookup2
Gender.lst:Gender:Person::Lookup2
letdate.lst:LetterDate:OtherDates::Lookup2
care.lst:hospital:service::Lookup2
```

Lists can be written in any text editor and are saved with an extension .lst. ExECT and IDEx gazetteers where created in Geany, https://www.geany.org/

Clinical terms mapping The original ExECT pipeline used UMLS (see Introduction, section 1.4.1) derived Bio-YODIE plugin for GATE, which identified biomedical terms and assigned to them an appropriate CUI. For the new version of the pipeline UMLS terms were used directly, without the Bio-YODIE plugin. Lists of diagnostic terms were extracted using the UMLS code browser implementation, https://github.com/arronlacey/UMLSBrowser[1] and converted to the GATE for-

[1] Developed for the project by Dr Arron Lacey, no longer maintained.

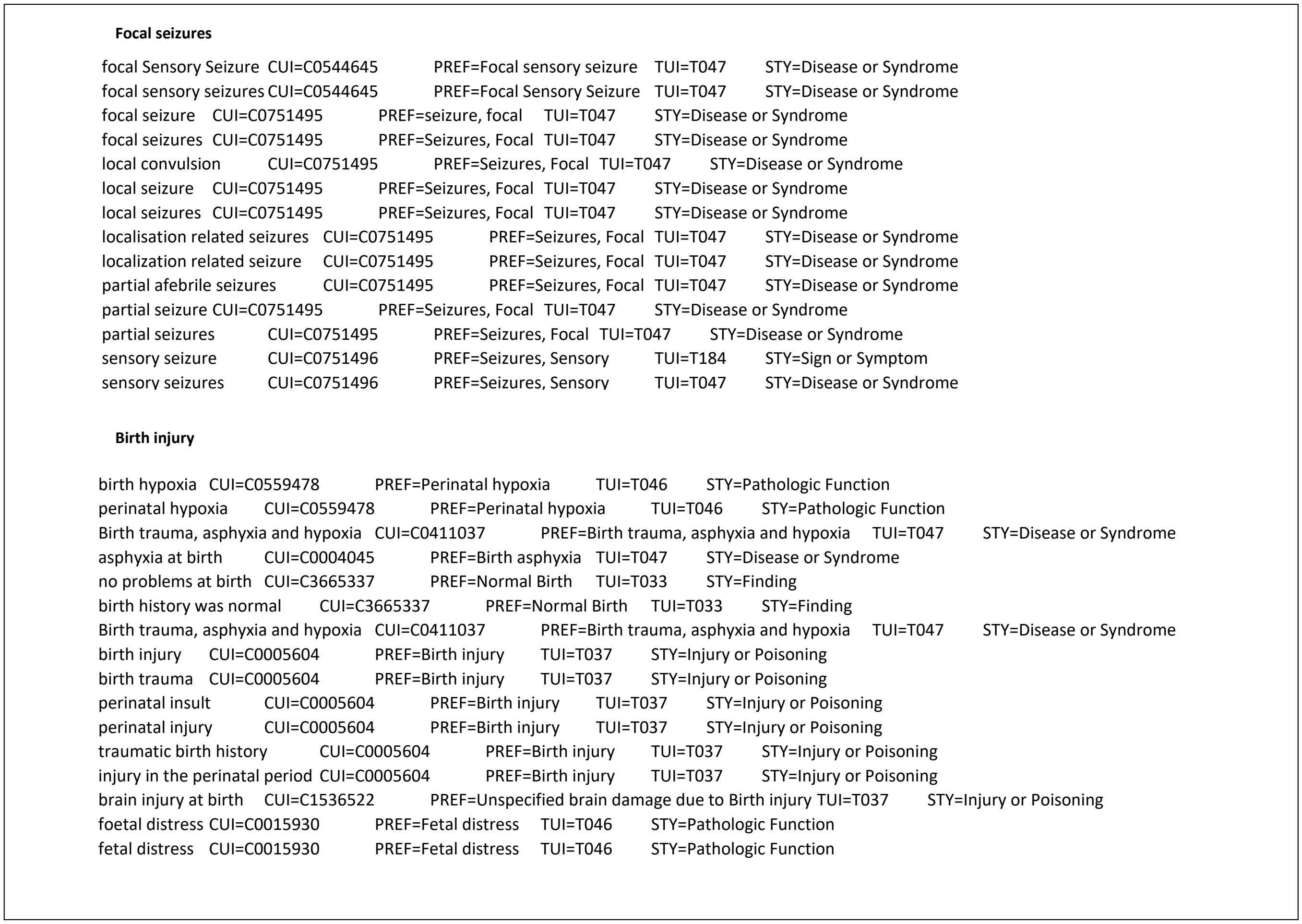
mat gazetteers. The browser was used to extract epilepsy (epilepsies, epilepsy syndromes, seizures), comorbidities, and events that may be associated with epilepsy, filtering for SNOMEDCT, HPO, MSH, and NCI vocabularies. The gazetteers were expanded by the addition of different forms of terms used, including English spelling and plurals. Throughout the testing phase new terms were added when needed. An example of entries for focal seizures from Seizure.lst and for birth injuries from Comorbidities.lst is shown in Fig.2.5. In addition to the phrase itself, it contains UMLS CUI and PREF (Preferred Term), TUI (Type Unique Identifier), and STY (Full Name Semantic Type).

2.4.3 ILAE classification

International League Against Epilepsy (ILAE) 2017 revised classification of seizures brought about some significant changes to seizure type terminology. [13] The classification is introduced in section 1.2.1
As the new terms were likely to begin to be more widely used, the gazetteers constructed were reviewed and some of the new terms were mapped to the existing concepts. For example, focal to bilateral tonic-clonic seizure was mapped to generalized tonic-clonic seizure with focal onset, which has the same CUI as secondarily generalised tonic-clonic seizure. Some concepts were not added such as focal onset myoclonic seizure as the term did not exist in the UNLS Metathesaurus at the time of the lists construction (it has been added since then).

2.4.4 Context implementation

The ConText algorithm for determining Negation, Experiencer, and Temporal Status from clinical reports [191] is available in GATE through the context implementation algorithm plugin. With the use of trigger terms and regular expressions it determines whether a Lookup refers to 'Patient' or 'Other', whether it is 'Affirmed' or 'Negated' and whether it is 'Recent', 'Historical', or 'Hypothetical'. The triggers may be located before a Lookup (pre-condition) or after it (post-condition). During the ExECT v2 development the effects of ConText algorithm had to be modified with rules for specific situation, for example reverse negation was used



```
Focal seizures

focal Sensory Seizure  CUI=C0544645	PREF=Focal sensory seizure	TUI=T047	STY=Disease or Syndrome
focal sensory seizures CUI=C0544645	PREF=Focal Sensory Seizure	TUI=T047	STY=Disease or Syndrome
focal seizure	CUI=C0751495	PREF=seizure, focal	TUI=T047	STY=Disease or Syndrome
focal seizures	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
local convulsion	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
local seizure	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
local seizures	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
localisation related seizures	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
localization related seizure	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
partial afebrile seizures	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
partial seizure CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
partial seizures	CUI=C0751495	PREF=Seizures, Focal	TUI=T047	STY=Disease or Syndrome
sensory seizure	CUI=C0751496	PREF=Seizures, Sensory	TUI=T184	STY=Sign or Symptom
sensory seizures	CUI=C0751496	PREF=Seizures, Sensory	TUI=T047	STY=Disease or Syndrome

Birth injury

birth hypoxia	CUI=C0559478	PREF=Perinatal hypoxia	TUI=T046	STY=Pathologic Function
perinatal hypoxia	CUI=C0559478	PREF=Perinatal hypoxia	TUI=T046	STY=Pathologic Function
Birth trauma, asphyxia and hypoxia	CUI=C0411037	PREF=Birth trauma, asphyxia and hypoxia	TUI=T047	STY=Disease or Syndrome
asphyxia at birth	CUI=C0004045	PREF=Birth asphyxia	TUI=T047	STY=Disease or Syndrome
no problems at birth	CUI=C3665337	PREF=Normal Birth	TUI=T033	STY=Finding
birth history was normal	CUI=C3665337	PREF=Normal Birth	TUI=T033	STY=Finding
Birth trauma, asphyxia and hypoxia	CUI=C0411037	PREF=Birth trauma, asphyxia and hypoxia	TUI=T047	STY=Disease or Syndrome
birth injury	CUI=C0005604	PREF=Birth injury	TUI=T037	STY=Injury or Poisoning
birth trauma	CUI=C0005604	PREF=Birth injury	TUI=T037	STY=Injury or Poisoning
perinatal insult	CUI=C0005604	PREF=Birth injury	TUI=T037	STY=Injury or Poisoning
perinatal injury	CUI=C0005604	PREF=Birth injury	TUI=T037	STY=Injury or Poisoning
traumatic birth history	CUI=C0005604	PREF=Birth injury	TUI=T037	STY=Injury or Poisoning
injury in the perinatal period CUI=C0005604	PREF=Birth injury	TUI=T037	STY=Injury or Poisoning
brain injury at birth	CUI=C1536522	PREF=Unspecified brain damage due to Birth injury TUI=T037	STY=Injury or Poisoning
foetal distress CUI=C0015930	PREF=Fetal distress	TUI=T046	STY=Pathologic Function
fetal distress	CUI=C0015930	PREF=Fetal distress	TUI=T046	STY=Pathologic Function
```

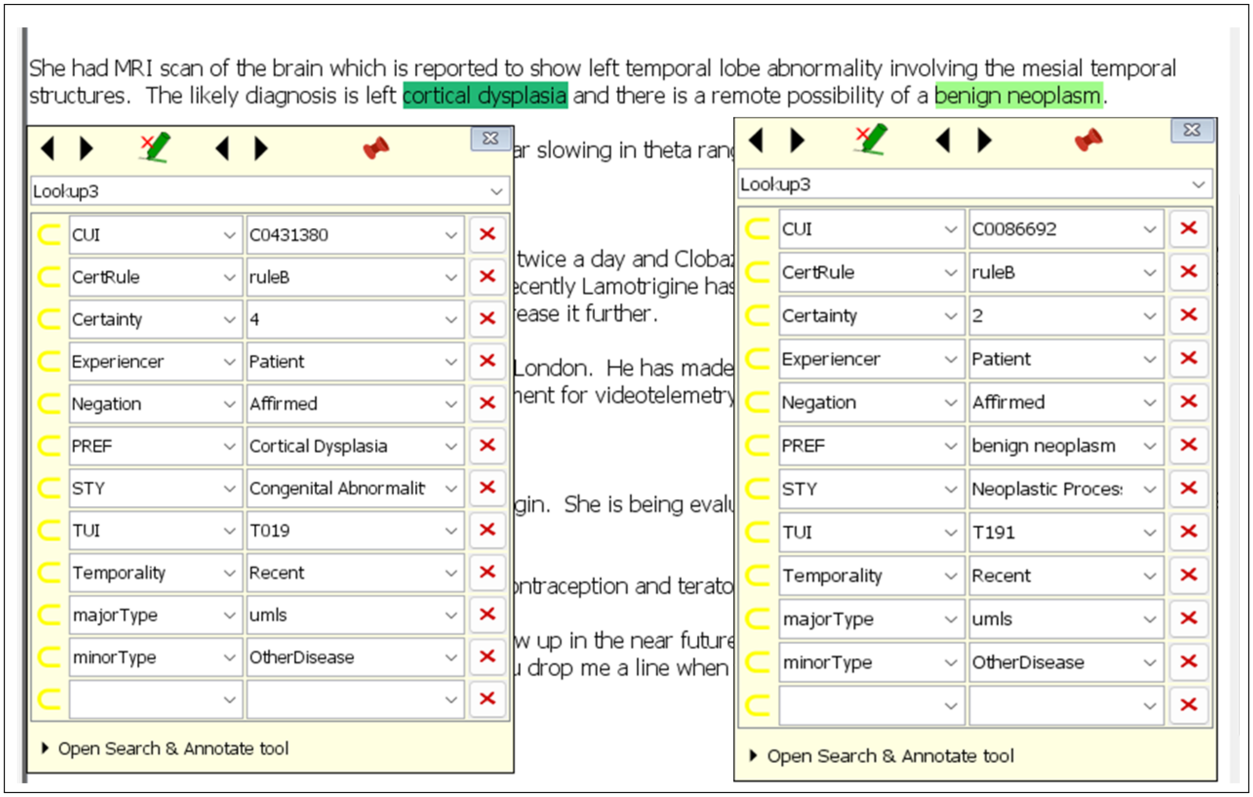
Figure 2.5: *The UMLS derived gazetteer structure with examples for entries relating to focal seizures (Seizures.lst), and birth injury (Comorbidities.lst) with the UMLS CUI, PREF, TUI, and STY separated by a tab.*

for terms such as non-epileptic attacks which is a term used to identify dissociative seizures and experiencer reversal rules were created when ‘Other’ was assigned incorrectly.

2.4.5 Certainty levels

Certainty levels were used in the original ExECT pipeline to quantify the level of certainty expressed in letters when a diagnosis of epilepsy or seizures was stated. Based on a list of terms expressing certainty and depending on the proximity of that term to a Lookup, not unlike ConText algorithm, a Lookup is annotated with a Certainty level, ranging from 1 for negative terms to 5 for definite statements. For example, in the sentence ‘*It is unlikely that his seizures are epileptic in nature.*’, the word ‘*seizures*’ would be assigned a certainty level of 2, whereas in ‘*The description of the episodes is consistent with temporal lobe epilepsy*’, ‘*epilepsy*’ would have a certainty level of 5. In ExECT v2 in addition to epilepsy and seizure phrases, Certainty was applied to all Comorbidities (Lookup3) with the additional rules written for lists of terms to ensure the Certainty levels tagging. The same terms were used in the annotation tests and in the gold standard set, with the list being included in the guidelines, Appendix B.2.

Fig. 2.6 shows Lookup annotation for two diagnostic phrases demonstrating ConText and Certainty application on the gazetteer terms, with the term ‘*likely*’ giving the certainty level 4 (left hand side annotation for Lookup 3) and the phrase ‘*remote possibility*’ resulting in the certainty level 2 (right hand side Lookup 3 annotation). Both Lookups are Affirmed, as there is no negation present, and both are of ‘Recent’ Temporality, with ‘Patient’ being identified as the Experiencer.



She had MRI scan of the brain which is reported to show left temporal lobe abnormality involving the mesial temporal structures. The likely diagnosis is left cortical dysplasia and there is a remote possibility of a benign neoplasm.

ar slowing in theta ran

twice a day and Cloba
ecently Lamotrigine has
ease it further.

London. He has made
ent for videotelemetry

gin. She is being evalu

ontraception and terato

w up in the near future
u drop me a line when

Lookup3

CUI	C0431380
CertRule	ruleB
Certainty	4
Experiencer	Patient
Negation	Affirmed
PREF	Cortical Dysplasia
STY	Congenital Abnormalit
TUI	T019
Temporality	Recent
majorType	umls
minorType	OtherDisease

Open Search & Annotate tool

Lookup3

CUI	C0086692
CertRule	ruleB
Certainty	2
Experiencer	Patient
Negation	Affirmed
PREF	benign neoplasm
STY	Neoplastic Proces:
TUI	T191
Temporality	Recent
majorType	umls
minorType	OtherDisease

Open Search & Annotate tool

Figure 2.6: *ExECT v2 Comorbidity lookup annotation with ConText and Certainty levels. The annotation on the left hand side shows Certainty level 4 triggered by 'likely', the anotation on the right gives the Certainty level 2 triggered by 'remote posibility'.*

2.4.6 Rule development

Java Annotation Pattern Engine (JAPE) was used to write rules for selecting and annotating the Lookups of interests. Using gazetteers, taggers, outputs from other rules, and a range of operators, a pattern to be matched was defined and the output to be produced specified. Operators allow to match the selected features of the Lookups or triggers by using '= =' as equal and '!=' as not equal, or to compare features using logical operators such as '<' , '<=' , '>=' , and '>'. For example, JAPE rule in Fig. 2.7 extracts onset date of a generalised tonic clonic seizures from the following statement '*She has infrequent generalised tonic clonic seizures. The first was during sleep in 2007 and last in 2008.*' It uses Person tagger, seizure

Lookup (seizure.lst), which has already been annotated with ConText to identify affirmed terms (‘Lookup.Negation == Affirmed’) and Experiencer, trigger phrases for onset from the Onset gazetteer, Certainty (default level of 5 as there are no Certainty triggers) and a rule based annotation for DateSince.

```
Rule: OnsetDateB4
Priority: 100

(
/* Lookcup from epilepsy.lst or seizure.lst, which is affirmed (not negated) and refers to
     a Patient (not Other) */
({Lookup.majorType == umls , Lookup.Negation == Affirmed, Lookup.minorType == Disease,
     Lookup.Experiencer == Patient}):item
/* end of sentence */
{Split}
/* onset term based on Onset.lst */
{Lookup2.language == onset , Lookup2.type == began}
/* any Lookup2 */
{Lookup2}
/* from a rule defining date since = term (since, in) and Date*/
{DateSince}
/* annotation output: Lookup with features as an Onset entity */
):match
-->
:item.Onset = {rule = OnsetDateB4, OnsetType = "date_of_onset",
/* UMLS CUI from the Lookup */
  CUI = :item.Lookup.CUI,
  PREF = :item.Lookup.PREF, TUI = :item.Lookup.TUI,
  STY = :item.Lookup.STY,  Negation = :item.Lookup.Negation, Experiencer = :item.Lookup.
     Experiencer,
/* Certainty level transferred from the  */
  Certainty = :item.Lookup.Certainty,
/* Date split into 3 elements, YearDate, MonthDate, DayDate */
  YearDate = :match.DateSince.YearDate,
  MonthDate =:match.DateSince.MonthDate,
  DayDate = :match.DateSince.DayDate}
```

Figure 2.7: *An example of JAPE rule for Onset annotation when the information of interest spans across two sentences.*

Standard annotation output

All JAPE rules follow the same patterns, with the Left-Hand Side (LHS) specifying the input (triggers, Lookups of selected types, outputs from other rules, operators) and the Right-Hand Side (RHS) that sets the annotation type and all the features to be assigned. Throughout the project only Lookups relating to

Patient, Affirmed (apart from febrile seizures), and of Recent temporality were extracted. All outputs were designed to have a common structure, with CUI, PREF, Certainty, and the rule name (for identification) as standard, with specific format for temporal expressions. For Age features the standard output is Age, AgeUnit, AgeLower, and AgeUpper, with the last two capturing age expressed as range, as in '*His seizures started at the age of 3 or 4*' (for Patient History output). Dates are annotated with three features: DayDate, MonthDate, and YearDate. For time period expressions such as in '*He suffered a stroke 2-3 years ago.*' the output contains TimePeriod, NumberOfTimePeriods, LowerNumberOfTimePeriods, and UpperNumberOfTimePeriods, with the last two features capturing the range of '*3 or 4*' in the example provided.

2.4.7 Groovy script

GATE provides a plugin, which facilitates the deployment of Groovy (Java-syntax-compatible object-oriented programming language) script https://groovy-lang.org/ within the GATE developer. This allows for the creation of scripts in the Groovy scripting console and running them as a PR on the documents. Groovy scripts were written to create output in CSV format of all annotations created by the pipeline, for the validation of ExECT v2 output and for the creation of datasets from the Epi25 cohort document set, for the SAIL genetic linkage study. Groovy PR for each annotation type was attached, in turn, to the end of the pipeline to produce an output file.

```

/* Creating an output file and identifying annotations to be extracted */
new File(scriptParams.outputFile).withWriterAppend{ out ->
/* Output annotation type: epilepsyCause */
  doc.getAnnotations("Output").get("EpilepsyCause").each{
    anno ->
    def f = anno.getFeatures()
/* listing features to be extracted */
    String[] id =  doc.getFeatures().get("gate.SourceURL").split("/")
    out.writeLine(/${id[-1]},${anno.start()},${anno.end()}, ${f.get('CUI')},"${f.get('PREF')}",/+
    /${f.get('Negation')},${f.get('Experiencer')},${f.get('Certainty')},${f.get('rule')},/)
  }
}
```

Figure 2.8: *Groovy script extracting Epilepsy Cause annotations produced by ExECT v2.*

2.4.8 ExECT v2 pipeline

The ExECT v2 pipeline was built within the GATE developer by placing the processing resources in sequence, as illustrated in Fig. 2.9. The final element of the extraction process depends on the analysis to be performed, and requires different levels of post-processing outside the pipeline using various data analysis tools.

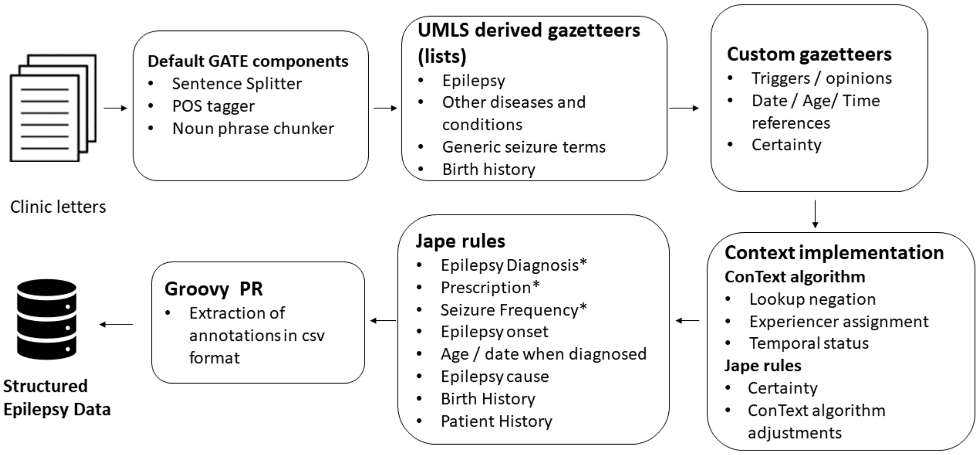


Figure 2.9: *Extraction of epilepsy clinical text (ExECT) v2 pipeline based on General architecture for text engineering (GATE) with custom epilepsy gazetteers, UMLS terms extract, ConText algorithms, Certainty level and Jape rules to create epilepsy annotations.*

2.5 Validation

The evaluation of the pipelines' performance was based on the measures derived from a contingency table, which compared the annotations made by ExECT v2 and IDEx to those from the gold standard sets, table 2.5. The method follows a modified version of evaluation metrics established by MUC-4 (the 4th conference on Message understanding) producing precision and recall, Eq.2.1 and Eq.2.2, which are then combined into a single F-measure, Eq.2.3. Precision and recall can be weighted in the calculation of F-measure, depending on the importance given to one over the other. [192]. Here they were given an equal weight, as shown in Eq.2.4.

	ExECT v2 and IDEx annotations	
Gold standard	**Pipeline annotated**	**Pipeline missed**
Annotated	True positive (TP)	False negative (FP)
Not annotated	False positive (FP)	True negative (FN)

Table 2.3: Confusion matrix for the evaluation of information extraction

$$Precision = \frac{TP}{TP - FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F = \frac{(1 + \beta^2) \times recall \times precision}{(\beta^2 \times precision) + recall} \tag{2.3}$$

Where β is a factor by which precision or recall can be weighted if either of them is considered to be more important. Otherwise it is set as 1 as in

$$F1 = 2 \times \frac{recall \times precision}{precision + recall} \tag{2.4}$$

Validation of the named entities in ExECT v2 and IDEx against the gold standard annotation set was carried out using the Corpus Quality Assurance tool within the GATE developer. This allows for the validation to be made on strict or lenient match and on different features. Strict match regards partially correct responses as spurious, hence overlapping annotations, but of different text span, would be considered erroneous. Lenient match allows for the overlapping annotations of different span to be included as correct. In the evaluation of the two pipelines' outputs, lenient match was applied with all features having to be correctly annotated (matched). Validation of any subgroups of the entities extracted, such as specific seizure types or comorbidities from Patient History annotations, and the Validation of IDEx against the SNB donors dataset was performed in R [193] using sqldf [194] and dplir. [195]. The latter was also used to perform per letter validation of the ExECT v2 annotation output and to compare the results produced to the original version of the pipeline.

Visual Studio Code

All Jape rules and groovy scripts were developed in Microsoft Visual Studio Code an open source code editor available at https://code.visualstudio.com/.

2.6 Genetic data linkage project

The aim of the *Genetic data linkage project* (Linking epilepsy next-generation sequencing datasets with routinely-collected healthcare records) was to link whole exome sequencing data to electronic healthcare records within the Secure Anonymised Information Linkage (SAIL) databank. Fig.2.10 shows the datasets and the processing needed to allow for the linked analysis to be performed within the SAIL Gateway.

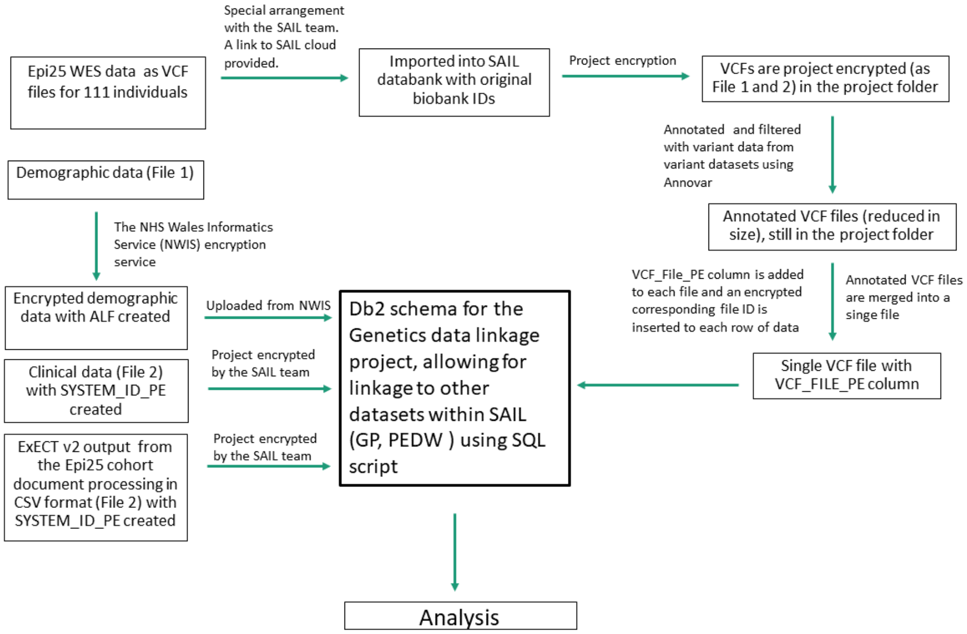

Epi25 WES data as VCF files for 111 individuals
Special arrangement with the SAIL team. A link to SAIL cloud provided.
Imported into SAIL databank with original biobank IDs
Project encryption
VCFs are project encrypted (as File 1 and 2) in the project folder
Annotated and filtered with variant data from variant datasets using Annovar
Annotated VCF files (reduced in size), still in the project folder
VCF_File_PE column is added to each file and an encrypted corresponding file ID is inserted to each row of data
Annotated VCF files are merged into a singe file
Single VCF file with VCF_FILE_PE column
Demographic data (File 1)
The NHS Wales Informatics Service (NWIS) encryption service
Encrypted demographic data with ALF created
Uploaded from NWIS
Clinical data (File 2) with SYSTEM_ID_PE created
Project encrypted by the SAIL team
ExECT v2 output from the Epi25 cohort document processing in CSV format (File 2) with SYSTEM_ID_PE created
Project encrypted by the SAIL team
Db2 schema for the Genetics data linkage project, allowing for linkage to other datasets within SAIL (GP, PEDW) using SQL script
Analysis

Figure 2.10: Genetic data linkage pipeline - uploading VCF files, clinical data from the SNB, ExECT v2 output, and individual demographic data for linkage with the routinely collected data within the SAIL databank

The SNB derived data for the project was uploaded into the SAIL databank following this approach. Personal Demographic data from the SNB Donors database formed File_1 and the Clinical dataset from the SNB Clinical database formed File_2. All ExECT v2 outputs created from the Epi25 document cohort were also uploaded as File_2 tables.

2.6.1 Epi25 Collaborative

Epi25 Collaborative (Epi25 Collaborative for Large-Scale Whole Genome Sequencing in Epilepsy) is an international collaborative project aiming to exome sequence up to 25,000 individuals with epilepsy to address some significant questions regarding the importance of rare and common variants, and *de novo* mutations in specific forms of epilepsy. [1].

The SNB has contributed annually to Epi25, providing DNA samples and phenotypic data for Genetic Generalised Epilepsy (GGE) and Non-Acquired Focal Epilepsy (NAFE). The submission process follows strict eligibility criteria for different epilepsy types. For GGE a convincing history of generalised seizures and generalised epileptiform on EEG, with normal neuroimaging is required, with history of focal seizures and moderate to profound intellectual disability being the exclusion criteria. For NAFE there needs to be a history of focal seizures, focal or normal EEG, and neuroimaging that is normal or showing hippocampal sclerosis, generalised seizures and moderate to profound intellectual disability being the exclusion criteria. From 2016 to 2018 the biobank contributed 169 samples to the Epi25 Collaborative, which then made the whole-exome sequence (WES) data available to the research team as Variant Call Format (VCF) files. These were utilised in the *Genetic data linkage project.*

2.6.2 VCF files upload into SAIL

VCF files relating to individuals who consented for their genetic data being used within the SAIL Databank were uploaded into the SAIL Gateway using a cloud link. Biobank numbers were used as file names which corresponded to System IDs assigned for Files 1 and 2. These were encrypted by the SAIL technical team following the upload.

2.6.3 Annovar

Annovar is a free variant annotating software available at http://www.openbioinformatics.org/annovar/. It can produce gene-based and filter-based annotations utilising specific databases. [196] For example, dbSNP, 1000 Genome Project, NHLBI-ESP 6500 exomes or Exome Aggregation Consortium

(ExAC) or Genome Aggregation Database (gnomAD), and find SIFT, PolyPhen, LRT, MutationTaster, Mutation Assessor, or CADD scores.

2.6.4 Annovar within the SAIL Gateway

Installing Annovar requires internet access and special arrangements were made with the SAIL technical team for the software to be installed. Individual annotation datasets were downloaded locally via Annovar website using Linux (Ubuntu) command line script:

```
-downdb -buildver hg19 -webfrom annovar refGene humandb/
```

and then uploaded to the SAIL gateway using ‘Request files in’procedure, and added to the allocated project directory. The following annotation datasets were uploaded: 1000g2014oct_eur, 1000g2014oct_all, clinvar_20200316, gnomad211_-genome, snp138, dbnsfp30a, and exac03.

Within the SAIL Gateway MobaXterm_Portable application (v20.2) (https://mobaxterm.mobatek.net/) was installed to access Linux command line to execute perl script.

2.6.5 Annotating VCF files

111 VCF files were firstly converted into ANNOVAR input format with the following command:

```
 ./convert2annovar.pl -format vcf4 biobank/file.vcf > file.avinput
```

This removed the metadata, producing .avinput text file, tab delimited columns containing chromosome, start position, end position, the reference nucleotides and the observed nucleotides, as shown in Fig.2.11

```

 1   565508    565508    G   A     . .
 1   567092    567092    T   C     . .
 1   752721    752721    A   G     . .
 1   756268    756268    G   A     . .
 1   830731    830731    T   C     . .
 1   838555    838555    C   A     . .
 1   840753    840753    T   C     . .
 1   846808    846808    C   T     . .
 1   861808    861808    A   G     . .
 1   866893    866893    T   C     . .
 1   868404    868404    C   T     . .
 1   881627    881627    G   A     . .
 1   888659    888659    T   C     . .
 1   894573    894573    G   A     . .
 1   897564    897564    T   C     . .
 1   900730    900730    G   A     . .
 1   903321    903321    G   A     . .
 1   918573    918573    A   G     . .
 1   919419    919419    T   C     . .
```

Figure 2.11: *VCF file after convert to ANNOVAR precessing with (L-R) Chr, Start, End, Ref, and Alt.*

Annotations were made using perl script shown below, with the datasets listed, producing .hg19_multianno file:

```
perl table_annovar.pl --buildver hg19 --out beata/file --remove --protocol refGene,1000
   g2014oct_eur,1000g2014oct_all,clinvar_20200316,gnomad211_genome,snp138,dbnsfp30a,exac03
    --operation gx,f,f,f,f,f,f,f --nastring '-' -otherinfo biobank/file.avinput  humandb/
```

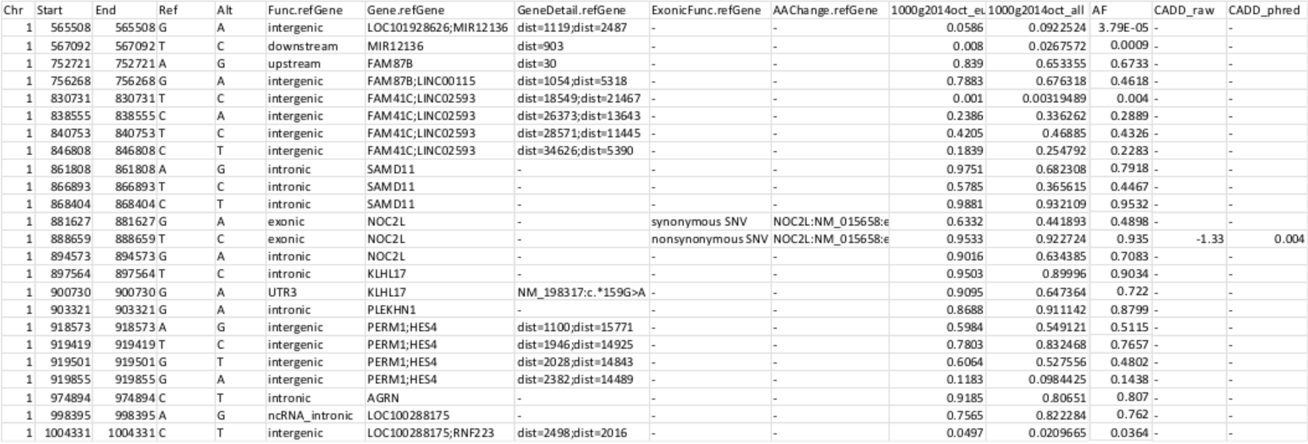
2.6.6 Crating variant data table

Following annotation a new column containing the encrypted VCF name (VCF_-FILE_PE) in each row of data was added to each text file which were merged into a single table containing all the annotated variants, Fig.2.12

```perl

# /usr/bin/env perl
# Run the script from within the directory with the txt files as follows:
# perl ../add_column_to_vcf.pl *.txt
# set the file extension to be used for backups
BEGIN { $^I = ".bak"; }
# set the warning switch to true
BEGIN { $^W = 1; }
# read the next line from the current file
while ( defined( $_ = readline ARGV ) ) {
# header lines start with Chromosome
  if ( /^Chr/ ) {
# append VCF_FILE_PE to header line
    s/$/\tVCF_FILE_PE/;
# get the number from the start from the file name
    ( $number ) = $ARGV =~ /^([0-9]+)/;
  } else {
# append current file number to row
    s/$/\t$number/;
  }
  print $_;
}
```

Figure 2.12: *Perl script creating a single VCF files with individual names added to the last column as VCF_FILE_PE*

The resulting dataset contained over 70 columns of annotations from different sources, allowing for various filtering during the analysis. Fig 2.13 shows a fragment of the annotated file with a selection of annotations used in the project, with their definitions provided below.



Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	1000g2014oct_eu	1000g2014oct_all	AF	CADD_raw	CADD_phred
1	565508	565508	G	A	intergenic	LOC101928626;MIR12136	dist=1119;dist=2487	-	-	0.0586	0.0922524	3.79E-05	-	-
1	567092	567092	T	C	downstream	MIR12136	dist=903	-	-	0.008	0.0267572	0.0009	-	-
1	752721	752721	A	G	upstream	FAM87B	dist=30	-	-	0.839	0.653355	0.6733	-	-
1	756268	756268	G	A	intergenic	FAM87B;LINC00115	dist=1054;dist=5318	-	-	0.7883	0.676318	0.4618	-	-
1	830731	830731	T	C	intergenic	FAM41C;LINC02593	dist=18549;dist=21467	-	-	0.001	0.00319489	0.004	-	-
1	838555	838555	C	A	intergenic	FAM41C;LINC02593	dist=26373;dist=13643	-	-	0.2386	0.336262	0.2889	-	-
1	840753	840753	T	C	intergenic	FAM41C;LINC02593	dist=28571;dist=11445	-	-	0.4205	0.46885	0.4326	-	-
1	846808	846808	C	T	intergenic	FAM41C;LINC02593	dist=34626;dist=5390	-	-	0.1839	0.254792	0.2283	-	-
1	861808	861808	A	G	intronic	SAMD11	-	-	-	0.9751	0.682308	0.7918	-	-
1	866893	866893	T	C	intronic	SAMD11	-	-	-	0.5785	0.365615	0.4467	-	-
1	868404	868404	C	T	intronic	SAMD11	-	-	-	0.9881	0.932109	0.9532	-	-
1	881627	881627	G	A	exonic	NOC2L	-	synonymous SNV	NOC2L:NM_015658:e	0.6332	0.441893	0.4898	-	-
1	888659	888659	T	C	exonic	NOC2L	-	nonsynonymous SNV	NOC2L:NM_015658:e	0.9533	0.922724	0.935	-1.33	0.004
1	894573	894573	G	A	intronic	NOC2L	-	-	-	0.9016	0.634385	0.7083	-	-
1	897564	897564	T	C	intronic	KLHL17	-	-	-	0.9503	0.89996	0.9034	-	-
1	900730	900730	G	A	UTR3	KLHL17	NM_198317:c.*159G>A	-	-	0.9095	0.647364	0.722	-	-
1	903321	903321	G	A	intronic	PLEKHN1	-	-	-	0.8688	0.911142	0.8799	-	-
1	918573	918573	A	G	intergenic	PERM1;HES4	dist=1100;dist=15771	-	-	0.5984	0.549121	0.5115	-	-
1	919419	919419	T	C	intergenic	PERM1;HES4	dist=1946;dist=14925	-	-	0.7803	0.832468	0.7657	-	-
1	919501	919501	G	T	intergenic	PERM1;HES4	dist=2028;dist=14843	-	-	0.6064	0.527556	0.4802	-	-
1	919855	919855	G	A	intergenic	PERM1;HES4	dist=2382;dist=14489	-	-	0.1183	0.0984425	0.1438	-	-
1	974894	974894	C	T	intronic	AGRN	-	-	-	0.9185	0.80651	0.807	-	-
1	998395	998395	A	G	ncRNA_intronic	LOC100288175	-	-	-	0.7565	0.822284	0.762	-	-
1	1004331	1004331	C	T	intergenic	LOC100288175;RNF223	dist=2498;dist=2016	-	-	0.0497	0.0209665	0.0364	-	-

Figure 2.13: *VCF file (fragment) containing Func.refGene (function of the gene affected), Gene.refGene (Gene name associated with one variant), GeneDetail.refGene (Distance from closest genes for intergenic variants), ExonicFunc.refGene (Exonic variant function (e.g., nonsynonymous, synonymous, frameshift insertion) AAChange.refGene (Amino acid change) 1000g2014oct_eur (variants present in the 1000 genomes 2014, European population) 1000g2014oct_all (variants present in the 1000 genomes 2014, All), AF(Allele frequency, global from gnomad211_genome), CADD_raw, and CADD_phred*

2.6.7 Allele frequency

Allele frequency (AF) annotations are derived from Genome Aggregation Database (gnomAD) v2.1., which includes 10,847 genomes from unrelated individuals sequenced as part of various disease-specific and population genetic studies, and reprocessed, for consistency, through the same pipeline. [197] Allele frequency or Alternate allele frequency, refers to the relative frequency of an allele at a particular locus in a population, and it is expressed as a proportion. [198] For this project, AF was used to select rare variants, which were defined as those with AF less than 0.001 and not occurring more often that twice within the cohort.

2.6.8 Combined Annotation-Dependent Depletion

Combined Annotation-Dependent Depletion were described in Chapter 1, section 1.7. CADD PHRED scores were used to select potentially damaging variants by filtering the scores that were ≥ 15 [199] together with the selected rare AF to identify rare and potentially damaging variants.

2.6.9 Extracting variants

Rare and damaging Variants were extracted using an SQL query that concatenated the values from Chr, Start, End, REF, Alt columns from the already filtered table with the specified AF and CADD values creating string identifying individual variants, as shown below:

```
CREATE TEMPORARY TABLE SAILw0661v.Epi25_AF_001_CADD_VAR  AS
SELECT CHR || '-' || 'START' || '-' || 'END' || '-' ||'REF'|| '-' ||ALT  AS VARIANT,
    GENE_REFGENE, EXONICFUNC_REFGENE, CADD_PHRED, VCF_FILE_PE
FROM SAILw0661v.EPI25_AF_001_CADD
```

These were further filtered for those not present more than twice in the study cohort and the resulting dataset was linked to other project datasets and analysed as shown in Fig.2.14.

2.6.10 SAIL data

The SAIL databank was described in section1.3.1 of the Introductory Chapter. The databank holds a broad range of routinely collected, anonymised health and other public service datasets for the Welsh population [200, 201]. Two SAIL datasets were used in the project, General Practice (GP) dataset and Patient Episode Database for Wales (PEDW)

GP dataset

GP practices maintain electronic health records for their patients, recording symptoms, test results, diagnosis, treatments and referrals to specialist services. Practices which have signed up to SAIL (85% of all GP practices within Wales in 2022) share these records through special software (Audit+) creating data extracts containing patient demographics (File_1), and clinical event details (File_-2). Within the SAIL Databank these are available as encrypted tables in the

SAILWLGPV (SAIL Welsh Longitudinal Practice Dataset) Schema as SAILWLGPV.PATIENT_ALF and SAILWLGPV.GP_EVENT. The former provides an encrypted patient identifier, ALF (Encrypted Anonymous Linking Field), WOB (week of birth) which is the date of the Monday that occurs prior to the actual date of birth, and is used as the date of birth in analysis, and GNDR_CD (Gender), coded as '1' for males, '2' for females, and '9' for unknown. The GP_EVENT table contains separate dated records for symptom, diagnosis, or treatment noted i.e. each row of data represents an individual event. Events are coded using the READ code system. GP datasets are refreshed quarterly in January, April, July, and October.

In the *Genetic data linkage project*, SAIL0661V.WLGP_PATIENT_ALF_-CLENSED_20200701 and SAIL0661V.WLGP_GP_EVENT_CLENSED_20200701 were used. Following SAIL file naming convention, 'SAIL0661' refers to the project number and ' CLENSED _2020701' indicates verified GP data up to 01.07.2020.

The tables were used to link individuals from the Epi25 cohort to GP data and ASM prescription (SQL script is given in Appendix D.1) READ code 'dn...'was used to capture ASM. [25]

Secondary care

The Patient Episode Database for Wales (PEDW) available within the SAIL databank is derived from an annual extract of diagnostic and treatment information relating to inpatient and daycase activity in Welsh hospitals and for Welsh residents treated elsewhere in the UK. The dataset is held within the SAILPEDWV schema and consists a number of tables that can be linked using different linking keys. SAILPEDWV_EPI provides information relating to an individual's episode of continues care of one consultant, including start/end of episode, length, and speciality. SAILPEDWV_SPELL provides data for all episodes within the SPELL and holds information on date of admission/discharge, method of admission/discharge, length of stay, and personal information (gender, geographical area of residence, and ALF). [202]. These tables can be joined and linked to SAILPEDWV.DIAG and SAILPEDWV.OPER to extract diagnostic and procedure information relating to the episode. SAILPEDWV.DIAG contains up to 14 diagnosis that could be assigned to a given hospital episode, according to priority e.g. primary, secondary etc. Diagnosis is coded using the International Classification of Diseases 10th Revision (ICD-10). SAILPEDWV.OPER gives the detail of operations performed within an episode, with potentially up to 12 procedures assigned in order i.e. primary, secondary etc. Procedures are coded using the Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures (OPCS4). [203]

In order to identify individuals with a history of unscheduled hospital admission with a diagnosis of epilepsy, SAILPEDWV.DIAG (SAIL0661V.PEDW _DIAG_-20200901) was used to extract diagnosis (G40) and linked to SAIL- PEDWV_-SPELL (SAIL0661V.PEDW_SPELL_20200901), identifying unscheduled hospital admissions (SQL script is given in Appendix D.1) .

2.6.11 ExECT data

The ExECT v2 output from the Epi25 cohort processing as CSV format files was produced using groovy script (2.4.7) for all annotation types. New letter reference, DOC, was created by substituting the biobank numbers within each letter reference with a random number, using Microsoft Excel RAND function, with the document sequence within each person's set being retained. The files were then uploaded as File_2 into the SAIL databank and added to the project schema following encryption, Fig.2.10

2.6.12 Linked data analysis

All linkage and analysis was performed within the SAIL gateway using Structured Query Language (SQL) for Db2 syntax queries run in Eclipse, and in R using sqldf package (https://CRAN.R-project.org/package=sqldf). All plots were produced with R ggplot. [204]

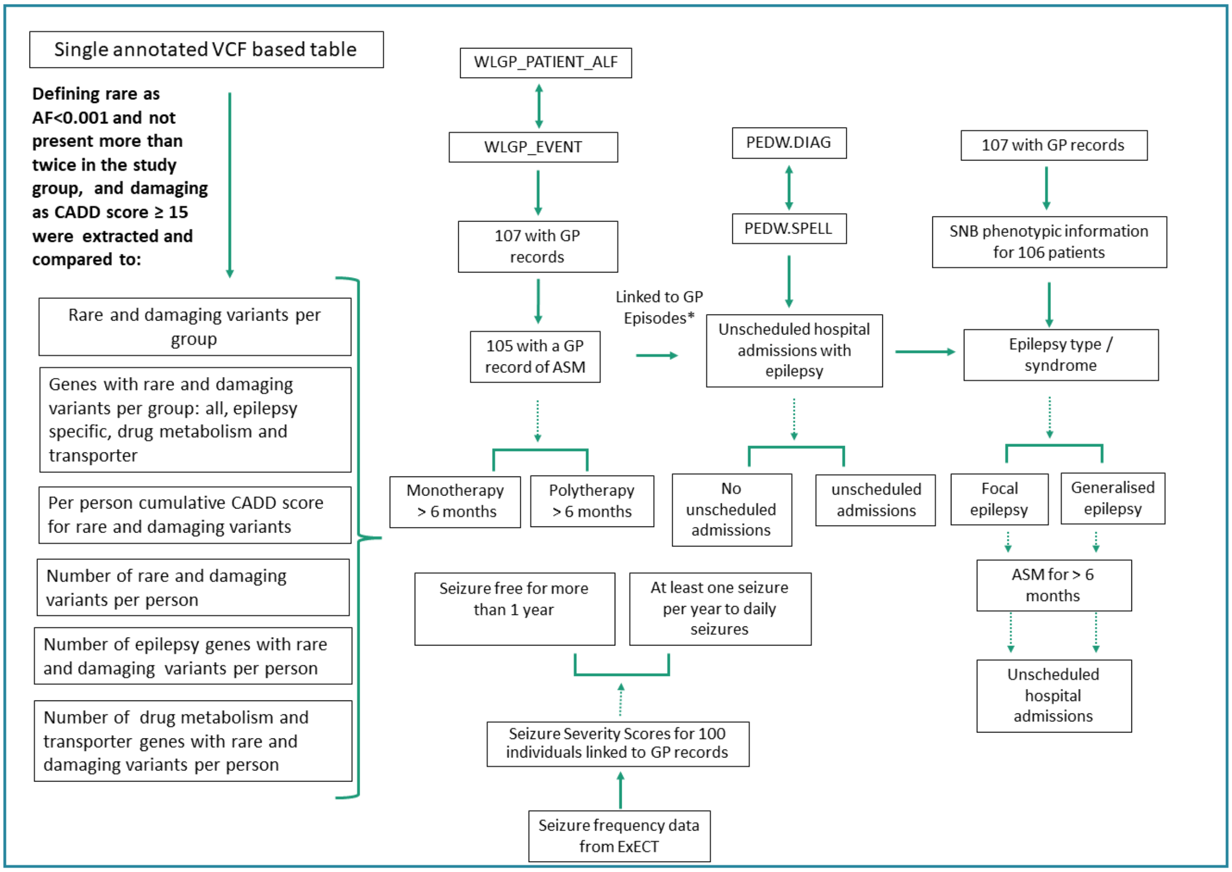
Datasets containing definitions and lists of terms needed for the analysis, e.g. genes associated with epilepsy, genes associated with drug metabolism and transportation, Residual Variation Intolerance Score (RVIS), and epilepsy/seizure lists, were added to the project schema using the standard SAIL document upload method.

Rare variants

For this project variants were defined as rare if they had an Allele frequency (AF) of <0.001 in the gnomAD exome collection (v2.1.1) and occurred not more than twice within the study group. Potentially damaging variants were those with Combined Annotation-Dependent Depletion (CADD) score ≥ 15 [205, 206].

Fig.2.14 shows the datasets linkage and analysis carried for the project with the results provided in Chapter 6 and some of the CL scripts used are given in Appendix D.1

.


Single annotated VCF based table
Defining rare as AF<0.001 and not present more than twice in the study group, and damaging as CADD score ≥ 15 were extracted and compared to:
Rare and damaging variants per group
Genes with rare and damaging variants per group: all, epilepsy specific, drug metabolism and transporter
Per person cumulative CADD score for rare and damaging variants
Number of rare and damaging variants per person
Number of epilepsy genes with rare and damaging variants per person
Number of drug metabolism and transporter genes with rare and damaging variants per person
WLGP_PATIENT_ALF
WLGP_EVENT
107 with GP records
105 with a GP record of ASM
Linked to GP Episodes*
Monotherapy > 6 months
Polytherapy > 6 months
PEDW.DIAG
PEDW.SPELL
Unscheduled hospital admissions with epilepsy
No unscheduled admissions
unscheduled admissions
107 with GP records
SNB phenotypic information for 106 patients
Epilepsy type / syndrome
Focal epilepsy
Generalised epilepsy
ASM for > 6 months
Unscheduled hospital admissions
Seizure free for more than 1 year
At least one seizure per year to daily seizures
Seizure Severity Scores for 100 individuals linked to GP records
Seizure frequency data from ExECT

* 99 had a diagnosis of epilepsy in GP records but all patients in the cohort had epilepsy as their diagnosis was confirmed before the Epi25 submission

Figure 2.14: *WES data, SNB dataset, and ExECT seizure frequency output linkage and analysis within the SAIL gateway*

2.6.13 Project Approval

All SAIL projects require approval of the Information governance review panel (IGRP). 'Linking epilepsy next-generation sequencing datasets with routinely-collected healthcare records' was approved by the IGRP in January 2019 with the project number 0661.

2.6.14 Version control

ExECT and IDEx development were version controlled using Gitkraken [207] for GitHub repository maintenance, https://github.com/swneurosci/ExECT-V2 for ExECT and https://github.com/BeataFS/IDEx for IDEx.

2.6.15 Literature search, library, and thesis writing

National Library of Medicine PubMed [208] and ScienceDirect [209], through the Swansea University institutional access, were used for literature searches and access. Library collection was maintained in Mendeley online (https://www.mendeley.com/) and Mendeley Desktop v1.19.8 with BibTeX syncing for exporting references in BibTeX format. The thesis was written using LaTeX2e in TeXstudio 4.2.1 [210]

2.7 Chapter summary

- All documents were sourced from the NHS hospital neurology clinics and preprocessed in different ways, including pseudonymisation for the development, test, and validation sets.
- The Swansea Neurobiology Biobank (SNB) was used to produce extracts from Donors and Clinical databases for the validation of IDEx, ExECT, and as File_1 and File_2 for the Genetic data linkage project.
- The gold standard development process helped to develop and finalise Annotation guidelines for epilepsy, which will be used in further projects. It also aided the development of Markup.
- General architecture for text engineering (GATE) was used to built the ExECT and IDEx pipelines providing an infrastructure for gazetteer and rule execution, and for creating structured annotations output with Groovy script.
- The involvement of the SNB in Epi25 Collaborative resulted in the 'Linking epilepsy next-generation sequencing datasets with routinely-collected healthcare records' project, a pathfinding research into linking genetic data to the health care datasets within the SAIL databank. It also provided an opportunity to link the

structured output produced by the NLP applications to the genetic data and the SAIL datasets within the gateway.

Chapter 3

GOLD STANDARD ANNOTATION

This chapter describes the construction of standard annotation specification for information extraction from epilepsy clinic letters and the creation of a gold standard annotation document-set. It gives the results of annotation tests performed and highlights the areas that were most challenging for the annotators involved in the process.

3.1 Standard annotation specification

The core elements of desired annotation outputs were established during the planning stage of ExECT v2. These were based on the original ExECT pipeline and the planned annotations, such as Patient History or Onset. Validation of ExECT v1 was carried out manually by a group of reviewers who compared and scored the extracted annotations against those made by a clinician. Although the annotations had to match fully to be judged as true positives, the precise format or phrasing was not critical; for example, dates could be expressed in different way as long as the meaning was the same, similarly, phrases used for epilepsy or seizure types did not have to be identical, as long as they represented the same diagnosis. In case of automated evaluation, such as offered by GATE's performance evaluation tools, or outside of the GATE developer when using data analysis tools, annotation features and formats have to be identical. The development of a standard annotation specification centred, therefore, not only on the entity and feature recognition but also reflected the format of the output produced by the ExECT v2 pipeline.

3.1.1 Test and validation sets

Clinic letters used in the validation set were derived from the SNB records of individuals who consented that their data can be used for epilepsy research, (see section 2.1.3) All letters were fully pseudonymised and de-identified, as described in section 2.1.1. This was done to preserve anonymity whilst maintaining the overall structure of the documents. Letters which contained information that appeared sensitive and potentially identifiable were replaced.

3.1.2 Markup configuration

A gold standard annotation set was developed using Markup(described in 2.3.2) by four annotators, including a consultant neurologist and three data scientists, two of whom worked on the redevelopment of the ExECT pipeline. The Markup configuration was created to specify the entities and features to be annotated. During annotation tests there were a number of changes to the configuration, all relating to the annotation features and their format; there were also many additions to the UMLS concept list. This process was also a test for Markup itself, as the application was still under development. With the group providing feedback and suggestions to the Markup designer, newer versions of the application were made available and used in the subsequent annotation sessions.

All changes to the configuration were recorded, with the final (9th), given in Appendix B.1, being used to create the validation set. Distinct entities of the configuration were: Diagnosis*, When Diagnosed, Onset, Epilepsy Cause, Seizure

*Attributes were developed by Huw Strafford

Frequency* Investigations*, Prescription*, Patient History, and Birth History. For each of these there was a set of attributes to be selected from a drop-down list.

All concept lists used in ExECT i.e. epilepsy, epileptic seizures, comorbidities, non-specific seizures (‘seizure slung’), and ASMs were combined into a single list of UMLS CUIs and PREF terms (see section 2.4.2).

3.1.3 Annotation guidelines for epilepsy

Annotation instructions were produced to assist the annotators, providing a comprehensive description of entities and attributes and guiding through the annotation process with numerous examples. They contained lists explaining the attributes to be assigned, such as investigation results, certainty phrases, or age expressions. As in the case of Markup configuration, the guidelines evolved and were refined during the annotation tests. Some entities and attributes were described in greater detail, more examples of difficult to annotate phrases and common mistakes were added, and the list of certainty phrases expanded. The final version of ‘What and how of annotating with Markup v7’which was accepted by the annotators and used later in further annotation tasks is given in Appendix B.2

3.2 Annotation tests

Before the main annotation task for the 100-letter validation set, experiments were carried out to assess the understanding of the process by the annotators. Four members of the team were each assigned a set of 10 pseudonymised clinic letters which they were to annotate using the guidelines provided for reference. These were previously unseen documents derived from the SNB records. Following the annotation process, letter by letter results were reviewed in a group session, where errors were analysed and an effort was made by the group to understand and agree on corrections. Any suggestions relating to the annotation guidelines were also discussed at this point and changes were implemented for the next test.

A new set of 10 clinic letters was allocated to the team members and the same process of review, error analysis, agreement on suggested corrections, and guidelines amendment followed. Results of the two ten-letter annotation experiments were compared to assess whether there were any changes in the level of agreement between the annotators.

A separate test was then performed by three annotators on 20 clinic letters to provide a benchmark for the later validation of the pipeline output against the gold standard.

3.2.1 Measuring agreement - Fleiss' Kappa

Inter-annotator agreement (IAA) is a standard way of assessing the reliability of an annotation process [180] and to measure the difficulty of the task. [**?**]. This is described in more detail in the Method chapter 2.2.2

For the 10-letter sets results IAA measured the difference in selecting a phrase under specific entity and not selecting it. It did not assess whether the selection was correct but simply measured the difference in the annotators choice. This was classified as 'YES' for selected phrase and 'No' when the same phrase was not selected. At this stage the presence or absence of features such as CUIs, certainty, or dates was ignored, although it was reviewed in group sessions. The results for each entity were amalgamated, however, as only single phrases relating to Birth History, Onset, When Diagnosed, and Epilepsy Cause were present in the sets, they were excluded from the IAA assessment.

Fleiss' Kappa, a variation of Cohen's Kappa for more than 2 raters [182], was used to calculate the IAA for both tests. It measures the agreement between annotators (raters) that is above that which may be attributed to chance. The general formula is given as

$$k = \frac{\bar{P} - \bar{P}e}{1 - \bar{P}e} \tag{3.1}$$

where P is the probability of agreement, Pe is the expected probability of agreement, P - Pe is the proportion of agreement obtained in excess of chance and 1 - Pe is the proportion of agreement that is attainable above chance, With k = 1 representing a full agreement and k $\leq$ 0 indicating no agreement. The results are shown in Table 3.1. Apart from Diagnosis, better agreement was observed for all annotation types in Test 2. Using interpretation provided by Landis and Koch [183], only Investigations showed 'substantial agreement', with Seizure Frequency and Prescriptions reaching 'moderate agreement'. In both tests IAA for Patient History was only 'slight'. For the scale used see section 2.2.2

		Test 1		**Test 2**		
	Annotations	Kappa	P-value	Annotations	Kappa	P-value
Diagnosis	31	0.408	0	20	0.097	0.286
Investigations	12	0.304	0.010	14	0.627	0
Patient History	35	0.001	0.984	30	0.166	0.026
Prescriptions	15	0.345	0.001	19	0.441	0
Seizure Frequency	21	0.159	0.075	9	0.429	0.002

Table 3.1: *Fleiss'Kappa, 10-letter Test 1 and Test 2, P-value signifies that the Fleiss' Kappa coefficient is statistically significantly different from 0 if p $\geq$ 0.5, Kappa of 0 would mean that there was no agreement.*

Annotation review

The most common issue for the annotators was the correct identification of phrases as entities, for example assigning non-specific seizures to Patient History rather than to the Diagnosis (epilepsy). It was also reported that the guidelines did not make it sufficiently clear what instances of seizures should be annotated.

There was significant uncertainty regarding comorbidities i.e. the group was not entirely clear which items should be included. This confusion was clearly reflected in the IAA recorded in both tests.

Diagnosis annotation was affected by the inclusion of non-specific seizures, but also by the instances where two types of seizures were present in a single annotation, as in *'intractable temporal lobe epilepsy'*, which some annotators treated as two diagnoses *'intractable epilepsy'* and *'temporal lobe epilepsy'* but others as one. The correct annotation (the former) was then agreed by the team and the instructions were amended accordingly.

Prescription annotations were affected by the inclusion of medication changes by some of the annotators, which was incorrect.

In the 10-letter set annotation task, Seizure Frequency was reported to be the most difficult to annotate in both tests, in terms of assigning features that identified events since or during a specific time period.

The two tests showed that the task of annotating clinic letters was difficult, required concentration and the ability to refrain from interpreting the information provided when it was ambiguous. These lessons were carried forward to the next annotation task which was set to create a benchmark for the validation results.

3.2.2 Measuring agreement - pairwise F measure

In the third and final test, 20 clinic letters were allocated to three annotators. The annotations were compared not only on named entity recognition but with respect to all features required for each entity. The resulting annotations were compared using GATE's Corpus Quality Assurance tool. The IAA plugin for uploading numerous annotation files (.ann) and a corresponding text file was used, Fig **??**. Validation was then performed on individual pairs of annotations in turn by selecting different reference (Key set) and response set. This method temporarily uses one annotator's set as a gold standard and another as a set that is being validated. [190] Precision, recall, and F1 score are then calculated using the formulae described in section 1.4.2.

Six sets of results were then exported and the average F1 score was calculated in Microsoft Excel for each Annotation type with macro/micro summary for the set a whole.

Entities extracted were: Diagnosis, Investigations, Onset, Patient History, Prescription, Seizure Frequency, When Diagnosed. There were no annotations to be made for Birth History or Epilepsy Cause, and this was confirmed in manual review of the letters. Features included in the validation were: Age, AgeLower, AgeUnit, AgeUpper, Certainty, CT_Performed, CT_Results, CUI, DayDate, DiagCategory, DoseUnit, DrugDose, DrugName, EEG_Performed, EEG_Results, EEG_Type, Frequency, FrequencyChange, LowerNumberOfSeizures, MonthDate, MRI_Performed, MRI_Results, Negation, NumberOfSeizures, NumberOfTimePeriods, PointInTime, TimePeriod, TimeSince_or_TimeOfEvent, UpperNumberOfSeizures, YearDate.

Annotation	**Average F1 score with features**	**Average F1 score without features**
Diagnosis	0.77	0.89
Investigations	0.83	0.89
Onset	0.57	0.69
Patient History	0.57	0.68
Prescription	0.79	0.98
Seizure Frequency	0.74	0.84
WhenDiagnosed	0.90	0.90
Macro summary	0.74	0.84
Micro summary	0.72	0.84

Table 3.2: *Pairwise average F1 score for IAA on 20-letter corpus with all features selected and without features*

Table 3.2 shows the results of the 20-letters test. ‘F1 scores with features’ shows annotations for which all attributes were matched. These scores can be considered as the benchmark for the validation of the ExECT v2 output (Chapter 3). It has to be noted that some of the scores were derived from a very small number of annotations, specifically Onset and When Diagnosed, with 6 or fewer phrases each. As an IAA for information retrieval it is a measure of choice, as unlike the Kappa statistic, it is not concerned with true negatives, which are in fact unknown. [186] The higher the F1 score, the greater the agreement among the annotators. For completeness and as the way of investigating the source of annotator disagreement, F1 for annotations without features was also calculated. That the scores were higher for all but one entity, suggests that at least in part the disagreement related to specific features. On reviewing the results, missing CUIs, differences in the certainty levels allocation, dose errors for prescriptions, and temporal concepts in seizure frequency were the most common errors. The same problem of missing or misallocating phrases to incorrect entities still occurred, and it was a reflection of the challenges presented by manual annotation.

3.3 Gold standard set for epilepsy

For the creation of a gold standard set, 100 clinic letters were allocated to four annotators, 25 each. Annotation guidelines, Markup configuration, and the UMLS concept list were updated to incorporate suggestions from the last test. Some of the letters included in the set were used previously but new blank .ann files were created. The annotated letters were reviewed by the whole team and if corrections were needed, they were approved by the entire group, for example missing CUIs, expressions of prescription frequency, or misallocation of entities.

There were discussions relating to spelling errors. Some felt that those not affecting the meaning of a phrases should be annotated, whereas others disagreed. Although it was clear that errors would affect the performance of ExECT's rules, which are based on dictionaries, it was decided that an error not leading to a significant ambiguity should still be annotated. For example, *'Generalised Tonic Chronic Seizure'* appeared a number of times in a letter, and as its meaning was very obvious, it was annotated, but not without a debate on the degree of misspellings that are acceptable. Similarly *'Hey diabetes '*, clearly means *'Her diabetes'* and was annotated as a comorbidity in Patient history, as *'her'* is one of the trigger for diagnostic terms, but this phrase would not be picked up by ExECT because of the error.

There were some statements in the letters that seemed more ambiguous than others and these were discussed and corrected i.e. the annotation was kept or deleted. For example, for the following statement annotated as onset *'She has been having frequent complex partial seizures for the last 1 year'* it was felt that there was no evidence that the seizures were not present before the period stated, i.e. that these were new events, and the annotation was removed from Onset. It was however kept for Seizure Frequency.

Overall it was felt that the quality of annotations was good and on correcting the errors which were identified, the set was kept as the gold standard against which the validation of the ExECT v2 could be carried out. It was agreed, however, that the task was difficult and that it was impossible to guarantee that some errors were not still present. This is something that was identified later during the validation process.

3.4 Chapter summary

- The creation of a standard annotation specification for epilepsy clinic letter using Markup was described in detail. The setting up of the common output structure was seen as the most important element for the development of a reference set of annotations for validating any annotation application.
- The process of development of annotation guidelines was described, including any changes suggested following the annotation tests. Provision of numerous examples of correct annotations and of common errors was judged to be very helpful by the annotators.
- The annotation guidelines were developed and expanded during the annotation tasks and will be a resource for further annotation projects. They are a good standard for annotation of epilepsy clinic letters.
- Annotation tasks were described and the two measures used for calculating IAA were reviewed. The pairwise IAA for the final set provided the benchmark to be used in the ExECT results validation

- The final annotation task that produced the gold standard set was outlined, including the process of review and error correction. A few examples of amended annotations were provided.

Chapter 4

ExECT v2 PIPELINE AND OUTPUT VALIDATION

This chapter describes the ExECT v2 pipeline output from the 100-letter gold standard set, a collection of structured data tables based on the entities annotated. It gives the results of the validation against the gold standard with the per mention and per letter scores. Examples of the most common errors are discussed. The performance of ExECT v2 is compared with ExECT v1.

4.1 ExECT v2 output

The process of ExECT v2 development, including discussion of all building blocks, is described in section 2.4, with the resulting pipeline shown in fig 2.9. The final step of the process involves creation of a set of data tables, one for each entity with the annotation features created by the JAPE rules becoming distinct data fields. This output is achieved by creating an additional processing resource (PR) containing a Groovy script for extracting each annotation type with features and adding it at the end on the pipeline. An example of the script extracting Patient History is shown in fig 4.1.

```

new File(scriptParams.outputFile).withWriterAppend{ out ->
  doc.getAnnotations("Output").get("PatientHistory").each{  -- Patient History
    anno ->
    def f = anno.getFeatures()
    String[] id =  doc.getFeatures().get("gate.SourceURL").split("/")
    -- START/END of annotation span and the UMLS CUI
    out.writeLine(/${id[-1]},${anno.start()},${anno.end()}, ${f.get('CUI')},/+
    -- Temporal features: time period and number of time periods from the event including
    ranges
    /"${f.get('PREF')}",${f.get('TimePeriod')}, ${f.get('NumberOfTimePeriods')},/+
    /${f.get('LowerNumberOfTimePeriods')},${f.get('UpperNumberOfTimePeriods')},/+
    -- dates
    / ${f.get('YearDate')},${f.get('MonthDate')},${f.get('DayDate')},${f.get('PointInTime
   ')},/+
    -- age including age axpressed as a range of upper and lower
    /${f.get('Age')},${f.get('AgeLower')},${f.get('AgeUpper')},${f.get('AgeUnit')},/+
    -- level of certainty , negation, and experiencer
    /${f.get('Certainty')},${f.get('Negation')},${f.get('Experiencer')},/+
    /${f.get('rule')}/)
  }
}

//Only kept "" for phrases which are likly to contain "," ; if "" are used for other
   features then the output is given in "" too


```

Figure 4.1: *Groovy script for the extraction of Patient History annotations from GATE developer listing all features to be extracted. The script is uploaded to the application as a processing resource (PR) and is added at the end of the pipeline. Output file destination and format have to be specified in the PR parameters before running the pipeline on a corpus of documents.*

Annotations produced by the application can be viewed and validated in the GATE developer using the built-in Corpus Quality Assurance tool. For '*per letter* ' results or subcategories of the main entities, for example specific diagnoses or items from patient history, validation has to be carried out outside the system, by analysing the extracted datasets using SQL, R, or other data analysis tools.
Table 4.1 lists annotations extracted by ExECT v2 with a short description of the features included, as more details are given when individual entities are validated. Common features of all annotations are: Annotation Start/End (these identify the location of an annotation within the document and help to distinguish between annotations with identical features within the same text), CUI, PREF (see section 2.4.2), Certainty, and Rule. Others are specific to the type of entity that is annotated with Onset, When Diagnosed, and Patient History producing annotations for Age/Time Since/Date.

Entity	Features
Birth History	Normal birth, birth injuries, gestational age as different categories of prematurity or term births
Diagnosis*	Epilepsy, epilepsy type, epilepsy syndrome, seizure type e.g. generalised tonic clonic seizure
Epilepsy Cause	Any events or diseases that are said to be the cause of an individual's epilepsy
Investigations	EEG, CT, MRI results annotated as normal or abnormal, but also whether they were performed
Onset	Age, time since, or date of first epileptic seizure (classified seizure) or epilepsy
Patient History	Any events, comorbidities that can be important in the context of epilepsy, unclassified seizures and common seizure terms (absence, fit) are also annotated with onset age/date if given, significant events such as traumatic head injury, brain tumours or surgery are annotated with age/date if provided
Prescription*	Antiseizure medication with quantity, unit, daily frequency, with only full prescription being extracted
Seizure frequency*	Generic seizure or specific seizure type and frequency expressed as number of seizures since or during specific time period, date or other point in time
When Diagnosed	Age, time since, or date of epilepsy diagnosis

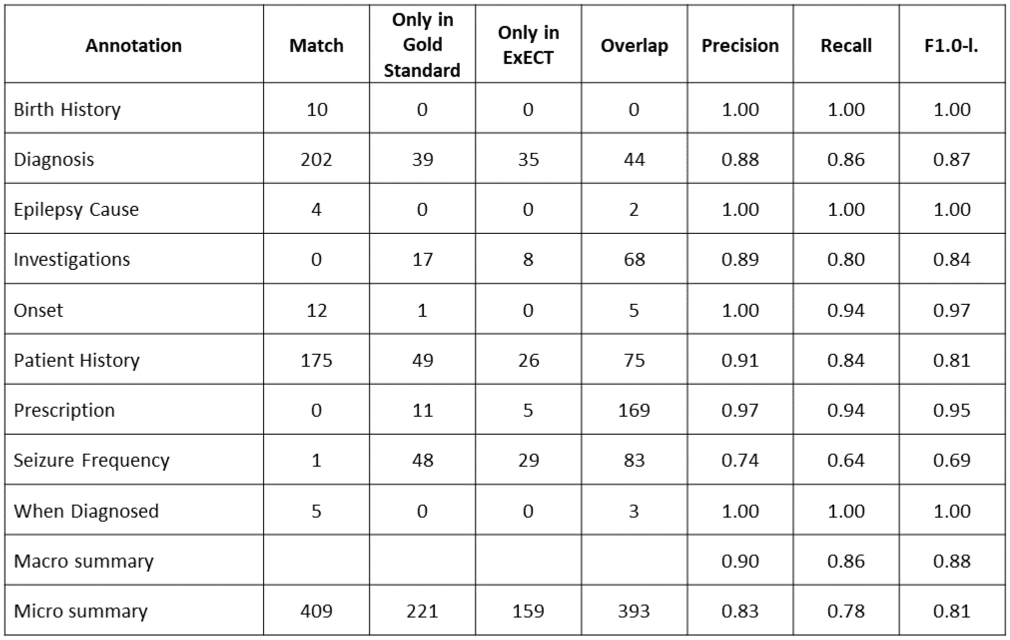
* Rules for these entities were redeveloped by Huw Strafford

Table 4.1: *Items (entities) extracted by the ExECT v2 pipeline and their features*

4.2 ExECT v2 Validation

4.2.1 Overall results – per item

Overall per item results comparing ExECT v2 and the gold standard annotations for the 100-letter corpus are provided by the GATE's Quality Assurance Processing Resource(PR) for the entire corpus, table. 4.2. The results are for all entities with all annotation features, except negation, which is extracted only for Febrile Seizures (as part of Patient History) and validated separately. Lenient F1-score is used in the validation which allows for the annotations of different lengths (covering longer or shorter text span but overlapping) to be matched. For example, prescription results are entirely based on overlapping annotations, as during the manual annotation of the validation set in Markup only the ASM name was highlighted, with features being selected from a drop-down list. In ExECT, JAPE rules select drug name, measurement, quantity, and dose in a single long annotation.



Annotation	Match	Only in Gold Standard	Only in ExECT	Overlap	Precision	Recall	F1.0-l.
Birth History	10	0	0	0	1.00	1.00	1.00
Diagnosis	202	39	35	44	0.88	0.86	0.87
Epilepsy Cause	4	0	0	2	1.00	1.00	1.00
Investigations	0	17	8	68	0.89	0.80	0.84
Onset	12	1	0	5	1.00	0.94	0.97
Patient History	175	49	26	75	0.91	0.84	0.81
Prescription	0	11	5	169	0.97	0.94	0.95
Seizure Frequency	1	48	29	83	0.74	0.64	0.69
When Diagnosed	5	0	0	3	1.00	1.00	1.00
Macro summary					0.90	0.86	0.88
Micro summary	409	221	159	393	0.83	0.78	0.81

Table 4.2: *ExECT v2 'per item' validation against the gold standard annotation set of 100 documents. Match = number of correct annotations with a strict match, Only in Gold Standard = number of annotations present only in the Gold Standard set (missing from ExECT), Only in ExECT = number of annotations present only in ExECT (spurious), Overlap = partial annotation, included in lenient match, Precision = correctly annotated items as a proportion of the number of annotations made by ExECT, Recall = number of correctly identified items as a proportion of all correct items, F1 = a weighted average of precision and recall, Micro summary = results for the entire corpus as one document, Macro summary = average results for precision, recall, and F1 score.*

4.2.2 Overall results – per letter

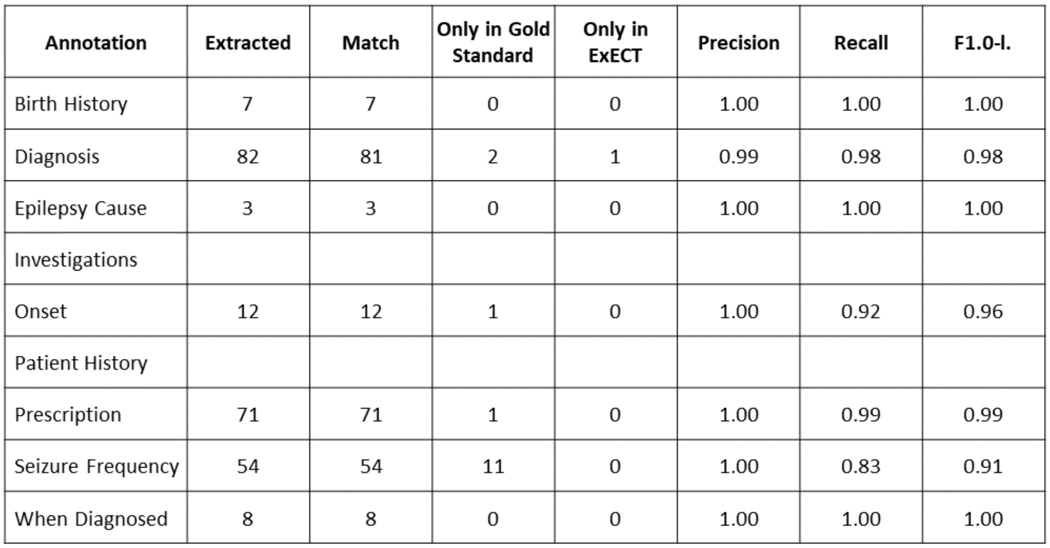
Per mention (item) results are a standard approach in evaluating the performance of an NLP information extraction system. [211] In a clinical setting a single dated record of investigation results, medication, or seizure frequency is sufficient for diagnostic or evaluation needs. For example, the current prescription may be mentioned a number of times in a letter, but these mentions may be viewed as a single record of the prescription fort that letter date. Per document score is, therefore, a valid method of assessing an application`s performance. [212] This approach also addresses the uneven distributions of each annotation type between the documents, similarly to the corpus macro summary shown in table. 4.2. Per letter scores are derived from at least one matching annotation of similar certainty i.e. levels 4 and 5 are treated equally as a positive outcome, as in the original ExECT pipeline validation. Most per letter results are based on more than one outcome. For annotations relating to grouped entities, such as Investigations (EEG, MRI, and CT) and Patient History, per letter results have no real value, as they do not indicate which items from the group have been extracted. For these, validation is performed separately, i.e. three elements of Investigations, and the most important or common variables in Patient History.

4.2.3 Results for specific entities

Birth History

10 Birth History annotations were extracted by ExECT v2, 4 related to birth injury (C0005604)*, 5 reporting normal birth (C3665337), and 1 an admission to special care baby unit (SCBU). Certainty levels were mostly 5 (8 cases). These were matched precisely against the gold standard annotations. Per letter results identified 7 documents with 7 true positive matches, 2 for birth injury, and 5 normal births.

*The numbers in () such as ‘C0005604‘’indicate CUIs.



Annotation	Extracted	Match	Only in Gold Standard	Only in ExECT	Precision	Recall	F1.0-l.
Birth History	7	7	0	0	1.00	1.00	1.00
Diagnosis	82	81	2	1	0.99	0.98	0.98
Epilepsy Cause	3	3	0	0	1.00	1.00	1.00
Investigations							
Onset	12	12	1	0	1.00	0.92	0.96
Patient History							
Prescription	71	71	1	0	1.00	0.99	0.99
Seizure Frequency	54	54	11	0	1.00	0.83	0.91
When Diagnosed	8	8	0	0	1.00	1.00	1.00

Table 4.3: *'Per letter' scores for ExECT v2 output against the gold standard annotation set of 100 documents. Investigations and Patient History are not included as they consist of distinct entities and are validated separately. Match represents strict and lenient matching, with all annotation features met and flexible approach to certainty, with level 4 and 5 treated as a positive result*

4.2.4 Diagnosis

Diagnosis annotations are designed to capture all positive mentions of epilepsy diagnosis, epilepsy syndromes, and specific seizure types. 281 such items were extracted by ExECT v2 and validated against 285 gold standard annotations. 246 items were matched (202 strict and 44 as overlaps) by LETTER, START, END, CUI, PREF, DIAGNOSTIC CATEGORY, and CERTAINTY, giving precision, recall, and F1 score of 0.88, 0.86, and 0.87 respectively.

On reviewing the 35 failed matches (false negatives and false positives) it appears that the most common reason for the errors were discrepancies in DIAGNOSTIC CATEGORY (missing or incorrect) and CERTAINTY (differences in the levels assigned). This could be seen when these features were removed from the corpus QA. Running the QA without DIAGNOSTIC CATEGORY feature selected increased precision, recall, and F1 score to 0.92, 0.91, and 0.92 respectively, and removing CERTAINTY produced identical results. When both features were removed, precision, recall, and F1 score increased to 0.96, 0.95, and 0.95.

The diagnostic categories are there to help differentiate between diagnosis derived from epilepsy, multiple seizures or a single seizure event, for example, an annotation of diagnosis from '*She has been diagnosed with temporal lobe epilepsy'*, results with the diagnostic category of '*epilepsy*' but '*She had a generalised tonic clonic seizure last week*' gives the diagnostic category of '*Single seizure*'.

The diagnostic categories have no great impact for an individuals with an existing diagnosis, certainty levels seem more significant as they indicate the level of diagnostic certainty. Yet they are more difficult to implement as triggers are used throughout the text and may affect items of interest when it is not desired. For example, in *'It may be that a number of his simple partial seizures have not been recognised.'*, the phrase*'may be'* is a certainty term (level 3) and in this context relates to the recognition of a number of seizures not to the seizures themselves. However, it affects the phrase *his simple partial seizures* giving it a certainty level of only 3, when in fact it should be 5, as there does not seem to be any doubt that the seizures are occurring.

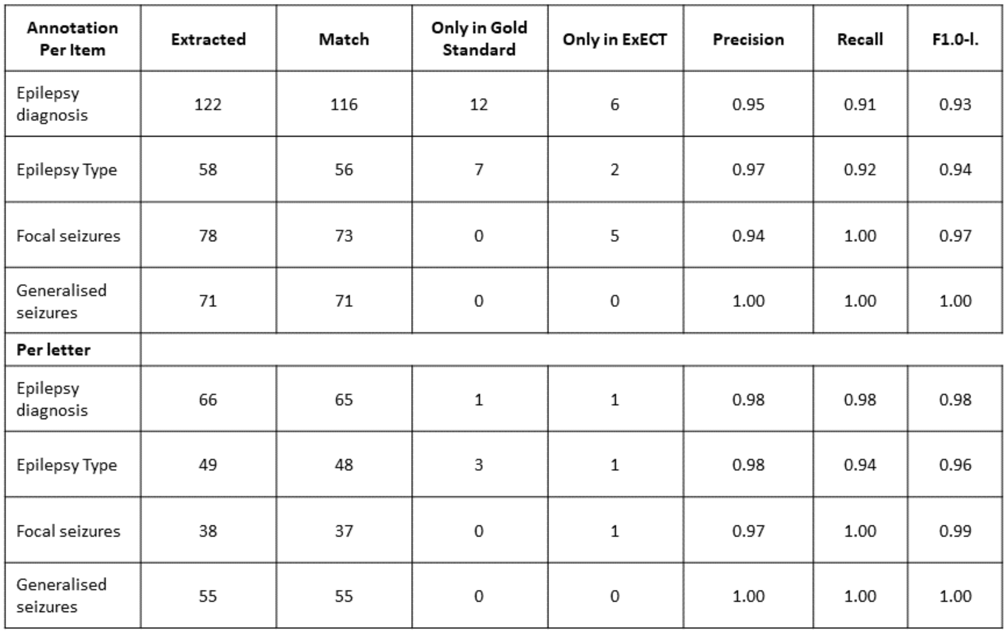
Other errors in the diagnosis extraction related to missed annotations, primarily, it seems, due to negation or experiencer context. For example: *'She came to the clinic accompanied by her son who has given me an eyewitness account of some of the episodes which are suggestive of epileptic seizures.'*. Her son is a trigger for 'Other'in the experiencer context implementation, which automatically assigns Lookups to 'Patient'or ' Other '. In this example an erroneously assigned Experiencer leads to the information being lost, as the rules are set to capture Lookups relating to patients only.

Epilepsy diagnostic terms are mentioned many times within each clinic letter, whether as epilepsy or seizures, hence per letter scores gave very high results. Out of 82 extracted diagnoses, 81 were matched against the gold standard annotations, with one false positive and 2 false negatives, giving precision, recall, and F1 score of 0.99, 0.98, and 0.98 respectively.

Diagnosis in ExECT v2 Vs ExECT v1

Diagnosis information extraction in ExECT v1 was different to that in the current application. Only the diagnostic statements consisted of epilepsy or epileptic seizures with a certainty level 4 or 5 were validated. Epilepsy type identification was assessed separately (focal and generalised) for annotations with a certainty level 4 or 5. Similar validation was performed for epileptic seizures, grouped as focal or generalised.

To carry out similar validation on the ExECT v2 diagnosis output, in csv format files from ExECT v2 and the gold standard annotation output (Markup) were analysed in R. Separate filtering for certainty levels 4 and 5 was performed, using DIAGNOSTIC CATEGORY to identify epilepsy-or seizure-based diagnosis. Linking the created subsets to Epilepsy and Seizure List allowed for the selection of CUIs by epilepsy type and, separately, by seizure type, and the extraction of annotations with epilepsy diagnosis or seizures. It is important to note that not all epilepsy or seizure types could be categorised. CUIs for epilepsy, symptomatic and refractory epilepsies were excluded from the analysis by epilepsy type, but were retained for the overall diagnosis of epilepsy. Nocturnal, tonic, and myoclonic seizures were not included in the seizure selection as they are not exclusive to either of the two categories used. The datasets were then compared and annotations representing true positives, false positives, and false negatives were identified to calculate F measures. Per item and per letter analyses were performed, and the results are given in table 4.4.



Annotation Per Item	Extracted	Match	Only in Gold Standard	Only in ExECT	Precision	Recall	F1.0-l.
Epilepsy diagnosis	122	116	12	6	0.95	0.91	0.93
Epilepsy Type	58	56	7	2	0.97	0.92	0.94
Focal seizures	78	73	0	5	0.94	1.00	0.97
Generalised seizures	71	71	0	0	1.00	1.00	1.00
Per letter							
Epilepsy diagnosis	66	65	1	1	0.98	0.98	0.98
Epilepsy Type	49	48	3	1	0.98	0.94	0.96
Focal seizures	38	37	0	1	0.97	1.00	0.99
Generalised seizures	55	55	0	0	1.00	1.00	1.00

Table 4.4: *ExECT v2, Extraction of epilepsy diagnosis categories with certainty level 4 or 5 validated against the gold standard set with all other annotation features, per item and per letter results*

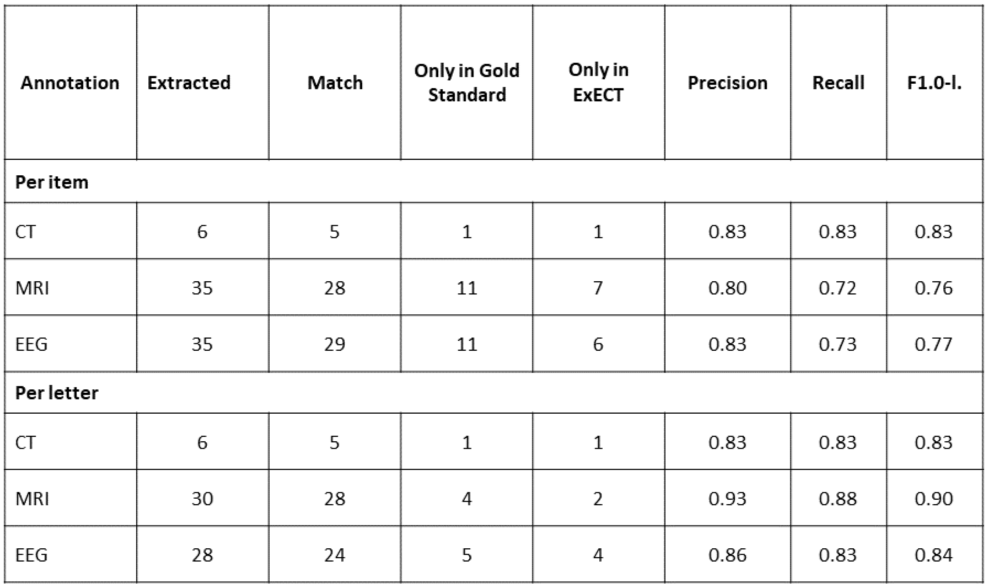
For all diagnosis subgroups, with the exception of focal seizure precision in per item scores, the performance of ExECT v2 appears better than that of ExECT v1. The most striking difference is in the extraction of generalised seizures. For the current version of the application, per item and per letter results were 1.00 for precision, recall, and F1-score. The corresponding results for ExECT v1 were, per item, 0.90, 0.52, 0.61, and per letter, 0.90, 0.68, 0.78, respectively. The closest results are for precision in focal seizure annotations, which for per item scores was lower for ExECT v2 than the in original pipeline at 0.94 compared to 0.96. The corresponding results for per letter scores were 0.97 and 0.96. However, with the effect of much lower recall, the F1 scores for both per item and per letter results for focal epilepsy were lower in the original pipeline than in the current version, with 0.81 and 0.89 compared to 0.97 and 0.99 respectively.

Epilepsy cause

Annotations of epilepsy cause are derived from very strict rules requiring a clear statement of causality, or one of many symptomatic epilepsy types being stated, which is, as other variables extracted, subjected to certainty levels. For this reason, but probably also due to the complex nature of assigning cause in epilepsy [213] the numbers of annotations are very small, 4 cases in per item and 3 in per letter sets. For the results produced, however, the validation scores against the gold standard annotations are the highest possible, with 1.00 for all measures.

Investigations

Investigation annotations extracted results for EEG, MRI, and CT scans mapped to abnormal/normal results using a list of specific abnormalities or terms expressing abnormal or normal. The output produced contained: LETTER, START, END, CUI, CT RESULT, EEG, EEG TYPE, EEG RESULT, MRI, MRI RESULT. 76 annotations were extracted giving 68 true positives, 8 false positives, and 17 false negatives. All annotations were matched as overlap, which illustrates one of the problems with extracting investigations information, where results are often given together for different tests, as is clearly visible in the following example: *'He had awake as well as sleep EEG which are reported to show irregular particularly during photic stimulation but no frank epileptic activity.'* where two EEGs (highlighted in yellow) relate to one result. Differences in the annotation spans at both ends make it harder to analysis the outputs for validation outside of GATE developer where using START/END of annotation is used to identify unique results.



Annotation	Extracted	Match	Only in Gold Standard	Only in ExECT	Precision	Recall	F1.0-l.
Per item							
CT	6	5	1	1	0.83	0.83	0.83
MRI	35	28	11	7	0.80	0.72	0.76
EEG	35	29	11	6	0.83	0.73	0.77
Per letter							
CT	6	5	1	1	0.83	0.83	0.83
MRI	30	28	4	2	0.93	0.88	0.90
EEG	28	24	5	4	0.86	0.83	0.84

Table 4.5: *ExECT v2 validation of CT, MRI, and EEG results against gold standard set, per item and per letter scores*

Most of the false negatives were caused by the absence of specific abnormalities reported or their expression in the results gazetteer. Given this, the overall scores for investigation extraction are not discouraging. In order to make a full evaluation, however, and to make a per letter assessment, investigation extraction had to be validated for each type of test separately using the csv format output from the ExECTv2 and gold standard annotations. These analyses were performed in R with the dplyr package, and the results are given in table 4.5. Similar scores are reached by CT, MRI, and EEG although the number of missed MRI and EEG results is the highest. The best per letter outcomes are produced by MRI result annotations, with F1 score of 0.90.

Investigations in ExECT v2 Vs ExECT v1

For the EEG results annotations, ExECT v2 produced slightly lower scores than the original pipeline for recall, 0.73 as compared to 0.75 for version 1, which led to a lower F1 score of 0.77 against 0.79. The results were similar for per letter validation, with precision, recall, and F1 score at 0.87, 0.90, and 0.88 respectively in ExECT v1, which was a little better than the current pipeline. For CT annotations, although it must be noted that the number of reports extracted was very small, there was an improvement in the application's performance as compared to ExECT v1, which reached precision, recall, and F1 score of 0.56, 0.59, and 0.57 per item, and 0.77, 0.63, and 0.69 per letter. Changes observed for the MRI results annotations are yet different, with the original pipeline performing better in per item extraction, at 0.82, 0.69, and 0.75 for precision, recall, and F1 score than ExECT v2, but slightly worse in the per letter assessment, at 0.87, 0.79, and 0.83 as compared to the current version, at 0.93, 0.88, and 0.90.

Onset

Onset is designed to capture information about an individual's age when they first experienced epileptic seizures. This information may not be provided in these terms in a clinic letter, hence the rules are set to extract any temporal data from which the individuals' age at onset can be extracted by linking to date of birth and clinic date. Only epilepsy and epileptic seizures are annotated, as non-specific seizures or events may not be epileptic in nature. These episodes are annotated separately by Patient History. Features extracted with Onset annotations include: LETTER, START, END, CUI, PREF, TIME PERIOD, NUMBER OF TIME PERIODS, LOWER NUMBER OF TIME PERIODS, UPPER NUMBER OF TIME PERIODS, YEAR DATE, MONTH DATE, DATE DATE, AGE, AGE UNIT, AGE LOWER, and AGE UPPER. Unlike diagnosis, onset is not mentioned that often in a clinic letter, it is usually discussed in first seizure clinic and may be reviewed when diagnosis is reassessed or mentioned in a referral to other services.

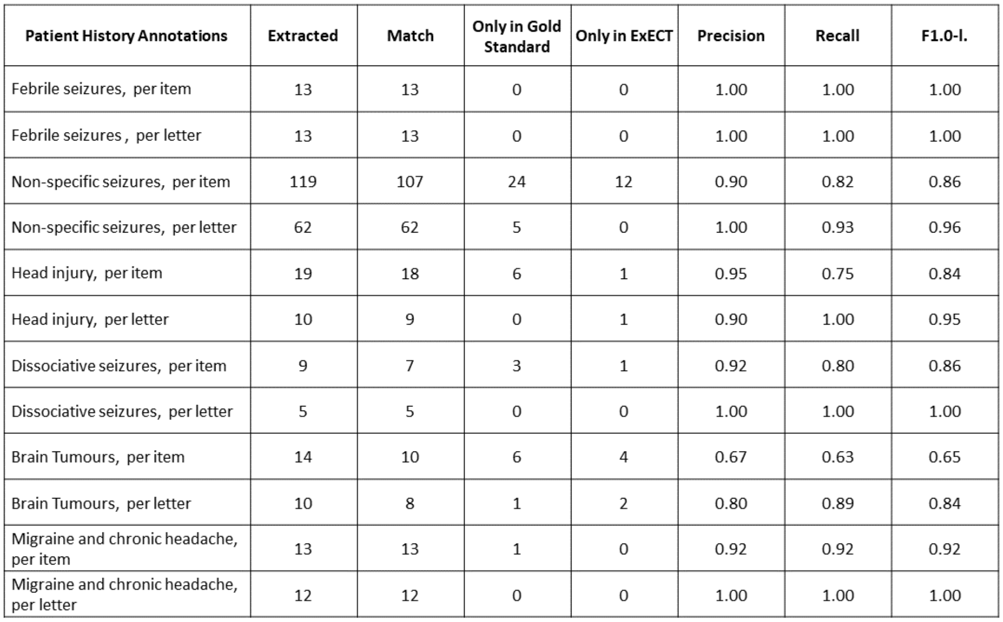
From the 100-letter set used in the validation, ExECT v2 extracted 17 records of onset, compared to 18 noted in the gold standard annotations. All of these were matched on all features (12 strict and 5 lenient matches), table 4.2. There was one false negative, which was for the generalised tonic clonic seizure in the following sentence: *'Her seizures however only started in 2007 when she suffered a generalised*

tonic clonic seizure.'. It appears that the generalised seizure was missed as, one the one hand, it was not close enough to the trigger word 'started ' and on the other, there was not another trigger, such as 'first' that would capture the event. Seizure (highlighted) is correctly captured by Patient History with the year date as indicated in the sentence, but this is not very helpful here.
For the per letter results in table 4.3, 12 true positives and 1 false negative were scored, with precision, recall, and F1 score at 1.00, 0.92, and 0.96.

Patient History

Results for Patient History shown in table 4.2 provide an overall score for a very wide category. Broadly speaking, Patient History contains annotations for events that are or may be associated with epilepsy, major comorbidities, non-specific seizures which may be epileptic seizures but were not clearly defined, dissociative seizures, and other seizure-like events. To carry out a thorough validation of Patient History annotations for these categories, they have to be evaluated separately using the csv format tables produced by the application. Output is the same for all Patient History annotations, except febrile seizures, for which NEGATION is added, and is similar to that extracted for Onset, with an additional temporal concept of POINT IN TIME. Table 4.6 shows the results per item and per letter for all categories validated. These were selected due to their importance for epilepsy diagnosis and treatment.

Febrile Seizures Febrile seizures are at present the only item for which positive and negative statements are annotated. Validation includes all features, although NEGATION which confirms a negative (Negated) or positive cases (Affirmed) is the most important. Certainty levels allocated in febrile seizures are supporting this dichotomy, with level 1 assigned to negative cases and levels 3, 4, 5 to positive cases. In this way Certainty is used somewhat differently than in other annotations. 13 febrile seizure annotations were extracted (CUI = C0009952), six affirmed (reporting history of febrile seizures) and seven negated (no history of febrile seizures). Four cases included age when seizures occurred, which was expressed as range (AGE LOWER, AGE UPPER, AGE UNIT). All annotations



Patient History Annotations	Extracted	Match	Only in Gold Standard	Only in ExECT	Precision	Recall	F1.0-l.
Febrile seizures, per item	13	13	0	0	1.00	1.00	1.00
Febrile seizures , per letter	13	13	0	0	1.00	1.00	1.00
Non-specific seizures, per item	119	107	24	12	0.90	0.82	0.86
Non-specific seizures, per letter	62	62	5	0	1.00	0.93	0.96
Head injury, per item	19	18	6	1	0.95	0.75	0.84
Head injury, per letter	10	9	0	1	0.90	1.00	0.95
Dissociative seizures, per item	9	7	3	1	0.92	0.80	0.86
Dissociative seizures, per letter	5	5	0	0	1.00	1.00	1.00
Brain Tumours, per item	14	10	6	4	0.67	0.63	0.65
Brain Tumours, per letter	10	8	1	2	0.80	0.89	0.84
Migraine and chronic headache, per item	13	13	1	0	0.92	0.92	0.92
Migraine and chronic headache, per letter	12	12	0	0	1.00	1.00	1.00

Table 4.6: *Patient History subcategories from ExECT v2 validation against gold standard annotation set, all features, with the addition of NEGATION for febrile seizures, Match = strict and lenient overlap and features.*

were correct as measured against the gold standard set, with no false negative or false positive cases, giving the F1 score of 1.00, for per item and per letter measure.

Non-specific seizures Seizures which are not fully defined, common seizure terms such as vacant episodes, absence, or fits, which were defined for the original ExECT pipeline as ‘seizure slang’ are annotated in Patient History. Some of these events may represent real epileptic seizures, for example absences in the context of an absence epilepsy diagnosis, or myoclonic jerks in Juvenile myoclonic epilepsy, and may be linked to Diagnosis output in post-processing. Without that context, however, they may represent non-specific seizure as the terms are used to mean many different events.

Non-specific seizures are the largest single component of the Patient History output, with 119 annotations extracted in ExECT v2 validation for seizures (C0036572), vacant episodes/absence (C0563606), and myoclonus/myoclonic jerks (C0027066). There were 107 true positive, 24 false negative, and 12 false positive cases. 13 annotations extracted contained features relating to onset, as age, date, or time since. Among these there was one false positive as shown in the following sentence: '*She was admitted following an unwitnessed seizure on 2nd October 2013*'. As no indication is given that this was a first seizure this annotation should not have contained temporal information, but it was, and it was not matched against the gold standard set. Out of the false negative cases, 2 contained onset features, as in (i)'*As you know she has experienced seizures since around 4 weeks after her operation.*'and (ii)'*Currently she get around 2-4 seizures per month. Although she did have a cluster of seizures in August, 2017*'. The first sentence is a clear statement of onset, however the second is more suggestive of seizure frequency and is a good example of the gold standard still containing annotations that may be questioned.

Overall results for non-specific seizure extraction are good, with F1 score of 0.88 per item and 0.96 per letter.

Head injuries Head injuries annotations were extracted for the following terms: head injury (C0497301), traumatic brain injury (C0870926), severe traumatic brain injury (C3508474), skull fracture (C0037304), and RTA (C0277693). Apart from being important to patient history, these annotations may be linked to onset and epilepsy cause using rules or through linkage to dated information in post-processing. Features extracted with head injury annotations are the same as for other Patient History items.

19 annotations were extracted by the pipeline, with 18 true positives giving a per item precision of 0.92. There was a relatively high number of false negatives, with gold standard identifying 6 cases missed by the pipeline. Here are some examples of the annotations missed, where the error was due to specific features not selected. '*His seizures started at the age of 3 or 4 he says shortly after he suffered a head injury in a road traffic accident.*', for which the 'head injury' was captured in ExECT without the age feature, but 'road traffic accident' has all

features required. This is due to the rule capturing the longest match, which when two items of interest follow the same age or date may cause the first item to be not fully annotated. It is not clear without reviewing the rules why this annotation missed the year date creating a false negative '*The following year in 2008 she had a fall down the stairs and suffered a traumatic brain injury.*' As references to a head injury may be repeated in a clinic letter, especially when it is seen as a possible cause of epilepsy and it is given prominence in the diagnosis information at the top of a letter, there are more opportunities for this information to be captured. This is probably the reason why per letter recall was increased, leading to a better F1 score of 0.95.

Dissociative seizures Dissociative seizures, more often referred to as non-epileptic seizures, are important to identify, especially for individuals who are treated for epilepsy, as these events may be mistaken for worsening seizures and lead to treatment changes that are not in fact needed. This issue is discussed in more detail in section 1.2.1 Two terms were used to extract dissociative seizure information from Patient History output, non-epileptic seizures (C3495874) and psychogenic seizures (C1142430).

9 annotations were extracted with 7 true and 2 false positive. There were 3 missed annotations. Precision, recall, and F1 score were, therefore, not very high, with 0.78, 0.70, and 0.74 respectively. On reviewing the errors, 1 false negative was due to certainty level 5 rather than 4, as in '*Dr X's impression was that these are non-epileptic attacks.*' where, 'impression' should assign the certainty of 4, but it failed. Another false negative was caused by the long distance between the subject '*He*' and the event '*a non-epileptic attack*' as shown in '*He was admitted to hospital and on reviewing the notes there are some atypical features including waxing and waning of motor activity and an observation form the neurology registrar on call was that it was more suggestive of a non-epileptic attack.*'. The person trigger that is needed to capture information in this sentences is far from the lookup, with a number of words, that could block the connection, for example '*reviewing, neurology registrar*' between the subject '*He*' and .the diagnostic term. Any adjustments

would be likely to make the rules too broad, resulting in a grater number of false positives.

The single false positive annotation was in the following statement: '*I am also passing on a copy of this letter to Dr Y, Consultant Psychiatrist to look into the possibility of non-epileptic seizures and do the needful.*', which is a hypothetical discussion and should not have been annotated, even with the Certainty level of 3, as it was.

When a possibility of dissociative seizures is being considered, the events in question are discussed and reviewed, resulting in many mentions throughout a clinic letter. This increases the possibility of capturing the information, as is clearly reflected in the per letter results, with precision, recall, and F1 score of 1 for each measure.

Brain tumours and surgery Extracting information on brain tumours and brain surgery is related to epilepsy cause, and the idea is, as with head injuries, that this information may be used in linked analysis. Although the list of terms for brain tumours and surgery contained in the Comorbidities gazetteer is longer, the following terms were identified as present in the two annotated sets (ExECT and gold standard) and used in annotation extraction and validation: gliosis (C0017639), mass lesion of temporal lobe (C2026258), brain neoplasms (C0006118), intracranial meningioma (C0349604), craniotomy (C0010280), hamartoma (C0349604), cerebral astrocytoma (C0750935), meningioma-surgery (C1096493), brain-astrocytoma (C3695127), and astrocytoma-surgery (C1096492).

14 annotations were extracted by ExECT v2 from the 100-letter set. 10 were fully matched against the gold standard, 4 were false positives and 6 were missed (false negatives). Per item results were precision at 0.67, recall at 0.63, and F1 score at 0.85, table 4.6. On reviewing these results, it is clear that 4 positive and 4 false negatives were brought about by erroneous allocation of Certainty. One example is shown here: *'The neuroradiology opinion is that it may be cortical*

dysplasia or a benign left temporal structural lesion' where the certainty level of 3 was applied only to cortical dysplasia but it should have also been assigned to temporal structural lesion. This is a common issue and rules (Certainty drift) have been deployed to address it, but in this case they appear not to have worked. One false positive case that was not due to certainty was in '*I have reassured her that I also can see no evidence of her having a brain tumour.*' for which the reverse negation rule caused the polarity error, as this is clearly a negated statement.

As references to brain tumours are likely to be made more than once in a clinic letter when tumours are diagnosed, there is a chance for per letter assessment to produce better results. This seems to be the case here, where out of 10 extracted letters, 8 were true positives, however the mismatch of certainty still produced 2 false negatives (letters with single mentions of tumours) and 1 false positive, the second being the case illustrated above. The final per letter scores were precision at 0.80, recall at 0.89, and F1 score at 0.84, which leaves some room for improvement.

Migraines and chronic headache The last category from patient history selected for separate evaluation was migraines and chronic headaches. These are comorbidities often reported by individuals with epilepsy and deserve special attention. Two terms were used in extracting annotations for this subgroup, migraine disorders (C0149931) and chronic headache (C0151293). There were 13 annotations produced by ExECT v2 and all matched against the gold standard annotations. There was one false negative, for which an existing rule does not seem to fire. As this was a single mention in a letter with 3 other correct annotations, this error did not carry across to the per letter results which were 1.00 for all performance measures.

Prescription

The validation of prescriptions is based on the full prescribing information containing ASM name, quantity, measurement (unit), and frequency. No certainty levels are applied. Per item validation was produced with the GATE's corpus QA plug-in and the results are shown in table. 4.2.

All 169 true positives were derived from lenient matches, as medication was annotated in Markup by highlighting drug name only, with drop-down lists for features, whereas in ExECT a whole phrase is selected. There were 5 false positives, and 11 false negatives, and these were investigated. Some of them were due to the presence of triggers that identify newly prescribed medications.
The approach taken in manual annotations and in ExECT was that new prescriptions initiated at the clinic could not be treated as the current prescription at the same clinic. There are trigger words that recognise such prescriptions and in same cases they affected the annotations. For example, in '*Recently Lamotrigine has been added and the present dose is 25mg in the morning and 50mg nocte with an intention to increase it further.*' was missed because of '*added*'. Similarly in '*He has been started on Levetiracetam 250mg once a day in a solution form.*' and in '*He has been re-started on Tegretol 200mg twice a day and Lamotrigine 50mg twice a day.*'It is likely that these prescription are affected by the same rules. Some errors, however, were caused by the annotations in the gold standard that did not follow the correct way of recording prescriptions as in '*She is currently taking Keppra 750 mg BD and folic acid 5 mg once-a-day.*', where the use of '*BD*' instead of '2' for frequency created both, a false positive and a false negative for ExECT. Despite these errors the overall results are very good, and compare well to the original ExECT pipeline which reported precision, recall, and F1 score at 0.96, 0.94, and 0.95 respectively.

To calculate per letter scores, outputs from ExECT and gold standard which consisted of LETTER, START, END, CUI, PREF, NAME, DOSE, UNIT, FREQUENCY were analysed in R. All brand names were converted to generic terms to allow for the correct extraction of the same drugs. There were two issues that proved hard to resolve in the analysis. Firstly, annotation START and END did not match in the two tables, because of the way the annotation was performed, which resulted in the absence of a unique annotation identifier which could be used for linkage (with LETTER id not being unique) and creation of duplicates. Secondly, full prescription for the per letter results consists of all ASM given with full doses and frequency. This was difficult to extrapolate since medication may be repeated a number of times in each clinic letter. Therefore, after the initial linkage and removal of duplicate values, records had to be inspected to identify full lists of ASM per letter. In the final calculation, there were 71 extracted records, 71 true positives, and 1 false negative, giving the final score for precision, recall, and F1 at 1.00, 0.99, and 0.99 respectively. These compare very well to the per letter scores produced by ExECT v1, which were 0.99 for precision, 0.91 for recall, and 0.95 for F1 score.

Seizure frequency

Seizure frequency is the most complex information to be extracted from clinic letters. There is no clear pattern as to how this information is recorded, proving quite a challenge for rule creation. It is also the most difficult task for the annotators who worked on the gold standard set.
Seizure Frequency annotates epileptic and non-specific seizures with features describing frequency changes, numbers of seizures, and temporal concepts. Output created by the pipeline includes: LETTER, START, END, CUI, FREQUENCY CHANGE, NUMBER OF SEIZURES, LOWER NUMBER OF SEIZURES, UPPER NUMBER OF SEIZURES, TIME PERIOD, NUMBER OF TIME PERIODS, LOWER NUMBER OF TIME PERIOD, UPPER NUMBER OF TIME PERIODS, SINCE/DURING, YEAR DATE, MONTH DATE, DAY DATE, and POINT IN TIME.

ExECT v2 extracted 113 seizure frequency annotations producing 84 true positives (1 strict and 83 lenient) and 29 false positives There were also 48 false negatives. Table 4.2 shows the full results. It is difficult to see a clear pattern in the errors. They seem to occur in the cases when frequency is not expressed as number of seizures per specified period of time but, as is the case in the three examples quoted here, when point in time is used or non-specific period of time. Fore example, '*She feels that they have been gradually worsening over the years* ', and '*She had a car crash a week ago and tells me she had a myoclonic jerk later that evening.*', and '*Her seizures remain reasonably controlled. She had 3 episodes in December, 3 in November and 10 in October.*'

For the false positive cases it seems somewhat similar as shown in the following, '*Since January John has taken his medication religiously and as you see above he has had far fewer seizures than ever before.*' and '*His main concern in clinic today was that he felt that he had some subjective short term memory loss. These has been occurring on a daily basis and is worsened following his admissions to hospital with seizures* ', in the second case memory loss episodes are mistaken for seizures.

For per letter results a similar process had to be followed as for prescriptions, because of the same problems with non-matching START and END of annotations. Frequencies for different seizure types were extracted to form a set for each letter. Non-specific seizures were ignored if specific seizures were present. With this approach, 54 letter annotations were extracted with 54 true positives; there were also 11 false negatives. The final result gave precision of 1, recall of 0.83, and F1 score at 0.91, which is a pronounced improvement as compared with the original pipeline scores of 0.92 for precision, 0.60 for recall, and 0.72 for F1 score.

When Diagnosed

When Diagnosed is the last annotation type extracted by ExECT v2. It is designed to catch very precise information on age, date, or time since epilepsy diagnosis. Output produced is the same as for Onset. As an illustration, the following sentence contains a trigger '*diagnosed*' which captures epilepsy type and age of diagnosis '*Mrs Jones has a past medical history of temporal lobe epilepsy diagnosed as a teenager and treated with Carbamazepine.*'Age expressed as '*teenager*' is converted to AGE LOWER = 13, AGE UPPER = 19, AGE UNIT = Year.

There were 8 annotations captured with no false positives or false negatives, producing a score of 1 for all measures. In per letter results, 8 true positives were extracted, with no false results.

4.3 Chapter summary

- This section described in detail the output produced by the ExECT v2 pipeline.
- The overall results of the validation performed against the gold standard set of 100 letters gave precision of 0.83, recall of 0.78, and F1 score of 0.81. Per letter scores were calculated for each entity, and most achieved F1 score of over 0.90, with a number of entities scoring 1.
- The validation results for specific diagnostic subgroups and investigations were compared to the original pipeline. Although some of the categories were difficult to compare, apart from the precision for focal seizures in the per item validation and the EEG results, the performance of ExECT v2 showed an improvident over ExECT v1. Similar results were observed for seizure frequency and prescriptions, with the most striking change in the per letter score for seizure frequency, from 0.72 in v1 to 0.91 in v2.
- The new entities of Birth History, Onset, Epilepsy Cause, and When Diagnosed produced very good results in the validation, they were however, based of a small number of annotations.
- A range of categories extracted under Patient History was discussed and validated. The best scores were achieved for febrile seizures (F1 score of 1) and the lowest for brain tumours (F1 score of 65) in the per item validation.

Chapter 5

SUPPLEMENTARY NLP PIPELINE FOR EXTRACTING IDENTIFIABLE INFORMATION - IDEx

This chapter describes a supplementary pipeline for the extraction of personal identifiable information from clinic letters, IDEx. Annotation entities and features are described and the validation of the first version of the pipeline, using a 200-letter set, is presented. Additional validation against the SNB donors' records is performed for the updated version of IDEx, extracting gender, Clinic, Letter, and Record date annotations from a random sample of 200 letters derived from the Epi25 cohort document set is also evaluated.

5.1 Extracting identifiable information from clinic letters

IDEx was designed as a separate pipeline for extracting personal demographic and identifiable information from clinic letters. It is a small application that can be deployed independently to produce a dataset for linkage, such as FILE 1 required for data uploads to the SAIL databank. [214]. See section 1.3.1 for details about the SAIL databank.

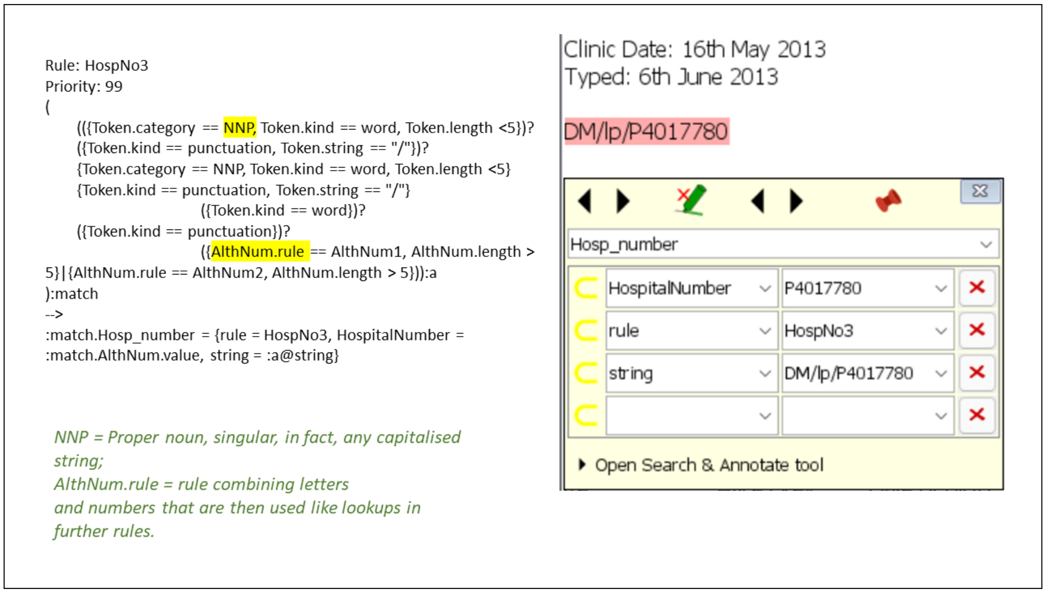
5.1.1 IDEx design and output

IDEx is based on GATE and uses the built-in elements, such as tokenisers and part of speech taggers, in addition to custom gazetteer lists and jape rules designed to capture specific entities. There is a single main gazetteer containing sixteen lists, four of which were imported from GATE's generic gazetteer (male/female names, and male/female titles) so that additional entries could be made, i.e. names or titles. Others contain clinic terms, reference phrases, hospital and NHS number phrases.

Development set

Rules were developed and tested on the 40-letter set of de-identified and pseudonymised documents used in the development of ExECT v1. Document preprocessing is described in section 2.1.1. A small number of additional letters from the biobank records was used to provide examples of different hospital number formats.

All JAPE rules were specifically designed for the application, apart from Clinic Date and Date of Birth, which were a modification of those written for ExECT v1. There are just under thirty rule sets for extracting six entities as shown in Table 5.1. Most rules rely on the presence of trigger phrases, some are based exclusively on context and a defined string pattern (Patient Postcode), others on pattern only, as illustrated in Figure 5.1. It shows one of the rules capturing Hospital Number, which is placed in a string containing, most likely, clinician's initials. In most cases Hospital Number follows a trigger, such as *'Hosp No:'* but it can also be given as part of a reference, which calls for rules that are very specific and most likely requiring adjustment for documents originating from different health boards, hospitals, or even departments.



```
Rule: HospNo3
Priority: 99
(
      (({Token.category == NNP, Token.kind == word, Token.length <5})?
      ({Token.kind == punctuation, Token.string == "/"})?
      {Token.category == NNP, Token.kind == word, Token.length <5}
      {Token.kind == punctuation, Token.string == "/"}
                          ({Token.kind == word})?
      ({Token.kind == punctuation})?
                          ({AlthNum.rule == AlthNum1, AlthNum.length >
5}|{AlthNum.rule == AlthNum2, AlthNum.length > 5})):a
):match
-->
:match.Hosp_number = {rule = HospNo3, HospitalNumber =
:match.AlthNum.value, string = :a@string}
```

NNP = Proper noun, singular, in fact, any capitalised string;
AlthNum.rule = rule combining letters and numbers that are then used like lookups in further rules.

Clinic Date: 16th May 2013
Typed: 6th June 2013

DM/lp/P4017780

Hosp_number

HospitalNumber	P4017780
rule	HospNo3
string	DM/lp/P4017780

Figure 5.1: *Hospital Number: JAPE rule to annotate specific pattern, using tokens and other rules (on the left) and Hosp_number annotation it produces in GATE developer, showing Hospital Number as one of the features (on the right)*

A fragment of JAPE script in Figure 5.2 illustrates two different types of rules for annotating Letter Date. The first one, more universal and robust *'Rule: LetDate'* uses a trigger phrase for letter dates based on a gazetteer. The second and third, *'LetDate2, LetDate3 '* are context dependent i.e. they rely on the presence of more general phrases (Lookups) that follow the letter date when the trigger is absent, and may be more sensitive.

```
Phase: Dates
# Input selects specific tokens from gazetteers to be used in the rules
Input: Lookup2 Lookup DateBio Person Hosp_number
Options: control=once  # once one matching annotation is found, stop annotating
Rule: LetDate
Priority: 100 # Priority given to the rule firing
# Date that follows a Letter date trigger and,  if there is a lookup between the terms,
  it mustn't be a  hospital term
(
({Lookup2.majorType == "LetterDate"}
({Lookup2.majorType != "hospital"})?
{DateBio}):a
):match
-->
:match.LetterDate = { rule = LetDate, value = :match.DateBio.value, string = :a@string,
  DayDate = :match.DateBio@DayDate,
  MonthDate = :match.DateBio@MonthDate,
  YearDate = :match.DateBio@YearDate, Date = :match.DateBio.Date}

# Date precedes a reference term such as 'Re:'
Rule: LetDate2
Priority: 99
(
({DateBio}):a
({Lookup2.majorType == "reference"})
):match
-->
:match.LetterDate = { rule = LetDate2, value = :match.DateBio.value, string = :a@string
 ,
  DayDate = :match.DateBio@DayDate, MonthDate = :match.DateBio@MonthDate, YearDate = :
 match.DateBio@YearDate,
  Date = :match.DateBio.Date}

# Date precedes a greeting  such as "Dear", if present,  and a title  i.e. "Dr"
Rule: LetDate3
Priority: 99
(
({DateBio}):a
({Lookup.majorType == "greeting"})?
({Lookup.majorType == "title"})

):match
-->
:match.LetterDate = { rule = LetDate3, value = :match.DateBio.value, string = :a@string
 ,
  DayDate = :match.DateBio@DayDate, MonthDate = :match.DateBio@MonthDate, YearDate = :
 match.DateBio@YearDate,
  Date = :match.DateBio.Date}
```

Figure 5.2: *JAPE script: rules for annotation of Letter Date (fragment)*

Entity	Detail
Clinic Date	Annotates clinic date converting any date format to features of Day Date, Month Date, and Year Date, in numeric format, triggered by clinic date terms, i.e. '*Clinic', 'Clinic date', 'Consultation date*'.
Letter Date	Letter date extracted in the same format as Clinic Date, triggered by context such as position in reference to other terms (Hospital Number, greeting e.g. 'Dear Dr ') and word triggers i.e. '*Date','Dictated','Typed* '.
Date of Birth	Annotates individual's date of birth in the same format as Clinic/Letter Date. Word triggers, variations of *'DoB'* are used and context i.e. person references, names.
Hospital Number	Annotates a combinations of numeric, alphanumeric, and punctuation tokens in a specified context i.e. presence of a health term or outside of it.
NHS Number	Only word triggers, such as *'NHS No'* are used to identify a 10-digit number, with and without spaces
Patient Postcode	Annotates a sequence of alphanumeric characters in a specified context, i.e. near patient name, date of birth, or NHS Number
Patient Gender	This is annotated, in the first instance, based on a title and name (using the gazetteers of male/female names) given in the reference section of a letter. If this fails, three consecutive personal pronouns are used to establish an individual's gender, as in *'She reports that her last seizure was at the weekend. She was tired after a night out with friends...'*.

Table 5.1: IDEx entities and feature description

5.2 Validation

IDEx was created and validated in stages. During the development of ExECT v1, rules for capturing Clinic Date and Date of Birth were written and validated manually. Following the construction of a separate IDEx pipeline the original 200-letter set was annotated with the new entities and features using BRAT annotation software. This set was used in a separate validation using GATE QA application.

At a later stage Patient Gender and Letter Date were added to the pipeline. The latter was introduced for cases when a Clinic Date was missing, so that a dated record could still be provided.

The extraction of Clinic and Letter Dates was evaluated on a random sample of 200 letters from the Epi25 cohort, whereas Patient Gender annotations were validated against the SNB dataset based on the Epi25 cohort letter set.

5.2.1 IDEx Validation on 200 epilepsy letters

Annotations using BRAT software were made for Date of Birth, Hospital Number, NHS Number, and Patient Postcode by a researcher not involved in IDEx algorithm construction. Annotations as .ann files were then uploaded into GATE developer and used to validate IDEx using the Quality Assurance Tool. IDEx rules are set to extract a single annotation of each entity, hence the figures shown in Table 5.2 are essentially per-letter scores, and show a perfect result.

Variables	Match	Overlap	Precision	Recall	F1 score
Date of Birth	199	0	1.00	1.00	1.00
Hospital Number	74	0	1.00	1.00	1.00
NHS Number	62	0	1.00	1.00	1.00
Patient Postcode	0	196	1.00	1.00	1.00
All	335	196	1.00	1.00	1.00

Table 5.2: *IDEx validation (partial), using the 200-letter set annotated using BRAT software.*

5.2.2 Validation of the Epi25 cohort output against the SNB Donors database

The SNB Donors database was used to provide personal demographic information about individuals from the Epi25 cohort to validate selected entities in IDEx pipeline. The structure and content of the SNB databases are described in Materials and methods, 2.1.3

Epi25 cohort IDEx extract

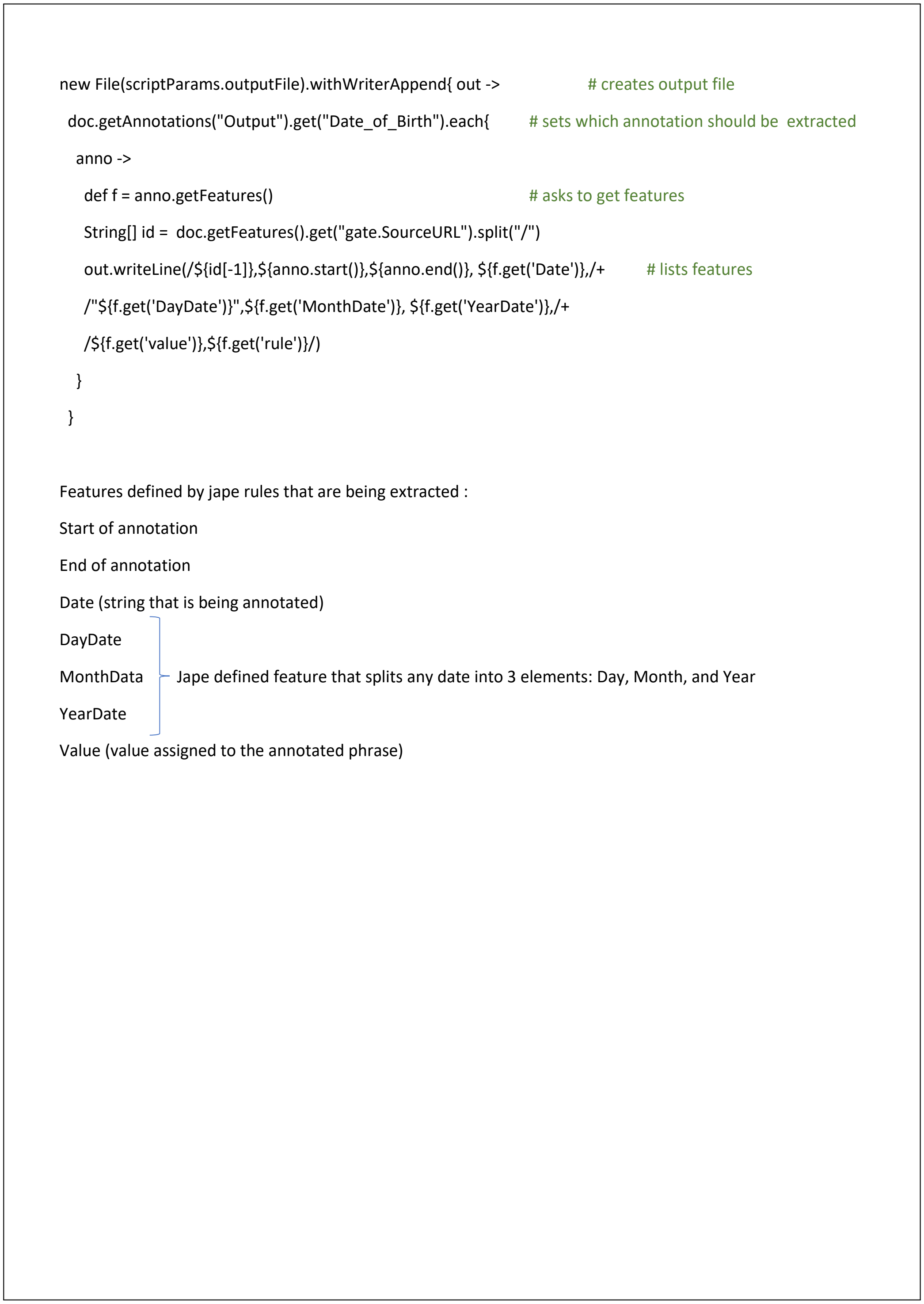
Epi25 cohort dataset is a set of authentic real life clinic documents for individuals who donated DNA samples to the SNB and were included in the SAIL genetic linkage study (Chapter 7). 771 documents relating to 111 individuals were annotated using IDEx. Annotations were extracted using Groovy script created for each annotation type and attached at the end of the pipeline. Figure 5.3 shows Groovy script for the extraction of Date of Birth and lists all features to be included.

IDEx output consisted of 6 data files: Epi25_ClinicDate, Epi25_LetterDate, DoB, NHS, Gender, and PostCode. Date fields were extracted to give a date stamp to the linked ExECT output and they were not being validated at this stage. Hospital numbers were not extracted as they could not be validated against the biobank records and they were not required for the ExECT output linkage.

Donors dataset

Biobank donors' details used in IDEx validation were extracted from the SNB Donors database using biobank numbers to identify individuals from the Epi25 cohort. A Microsoft Access query was used to extract NHS number, Date of Birth, Post Code, and Gender, with the results exported in a csv format file. This file was also used as FILE 1 for the SAIL genetic data linkage study (Chapter 7).

IDEx validation was carried out using sqldf package in R Studio. Figure 5.4 gives a fragment of the script used.



```
new File(scriptParams.outputFile).withWriterAppend{ out ->              # creates output file
 doc.getAnnotations("Output").get("Date_of_Birth").each{        # sets which annotation should be  extracted
  anno ->
   def f = anno.getFeatures()                                   # asks to get features
   String[] id =  doc.getFeatures().get("gate.SourceURL").split("/")
   out.writeLine(/${id[-1]},${anno.start()},${anno.end()}, ${f.get('Date')},/+        # lists features
   /"${f.get('DayDate')}",${f.get('MonthDate')}, ${f.get('YearDate')},/+
   /${f.get('value')},${f.get('rule')}/)
  }
 }
```

Features defined by jape rules that are being extracted :

Start of annotation

End of annotation

Date (string that is being annotated)

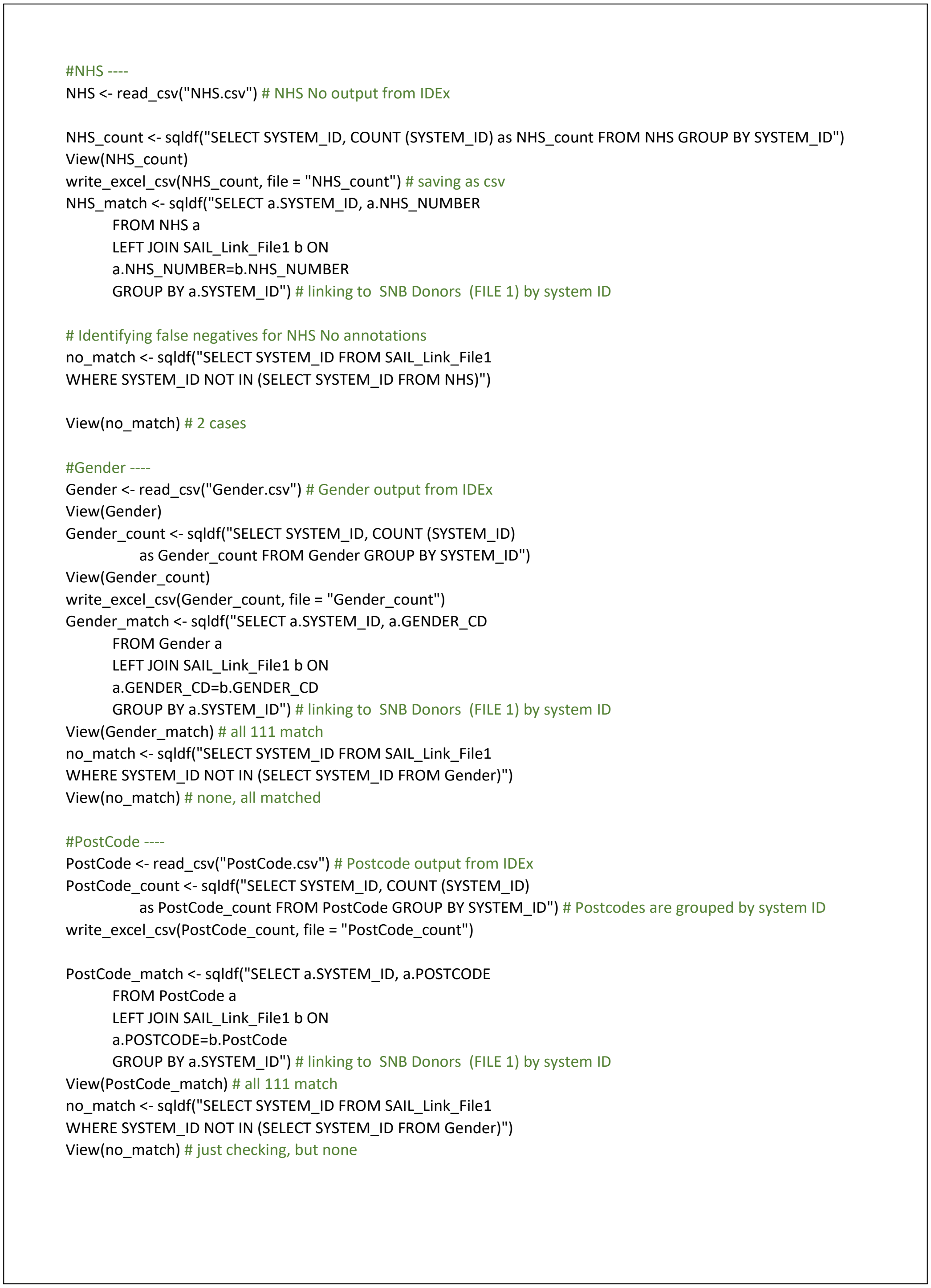
DayDate

MonthData — Jape defined feature that splits any date into 3 elements: Day, Month, and Year

YearDate

Value (value assigned to the annotated phrase)

Figure 5.3: Groovy script for Date of Birth annotation output and the list of features extracted



```
#NHS ----
NHS <- read_csv("NHS.csv") # NHS No output from IDEx

NHS_count <- sqldf("SELECT SYSTEM_ID, COUNT (SYSTEM_ID) as NHS_count FROM NHS GROUP BY SYSTEM_ID")
View(NHS_count)
write_excel_csv(NHS_count, file = "NHS_count") # saving as csv
NHS_match <- sqldf("SELECT a.SYSTEM_ID, a.NHS_NUMBER
        FROM NHS a
        LEFT JOIN SAIL_Link_File1 b ON
        a.NHS_NUMBER=b.NHS_NUMBER
        GROUP BY a.SYSTEM_ID") # linking to  SNB Donors  (FILE 1) by system ID

# Identifying false negatives for NHS No annotations
no_match <- sqldf("SELECT SYSTEM_ID FROM SAIL_Link_File1
WHERE SYSTEM_ID NOT IN (SELECT SYSTEM_ID FROM NHS)")

View(no_match) # 2 cases

#Gender ----
Gender <- read_csv("Gender.csv") # Gender output from IDEx
View(Gender)
Gender_count <- sqldf("SELECT SYSTEM_ID, COUNT (SYSTEM_ID)
            as Gender_count FROM Gender GROUP BY SYSTEM_ID")
View(Gender_count)
write_excel_csv(Gender_count, file = "Gender_count")
Gender_match <- sqldf("SELECT a.SYSTEM_ID, a.GENDER_CD
        FROM Gender a
        LEFT JOIN SAIL_Link_File1 b ON
        a.GENDER_CD=b.GENDER_CD
        GROUP BY a.SYSTEM_ID") # linking to  SNB Donors  (FILE 1) by system ID
View(Gender_match) # all 111 match
no_match <- sqldf("SELECT SYSTEM_ID FROM SAIL_Link_File1
WHERE SYSTEM_ID NOT IN (SELECT SYSTEM_ID FROM Gender)")
View(no_match) # none, all matched

#PostCode ----
PostCode <- read_csv("PostCode.csv") # Postcode output from IDEx
PostCode_count <- sqldf("SELECT SYSTEM_ID, COUNT (SYSTEM_ID)
            as PostCode_count FROM PostCode GROUP BY SYSTEM_ID") # Postcodes are grouped by system ID
write_excel_csv(PostCode_count, file = "PostCode_count")

PostCode_match <- sqldf("SELECT a.SYSTEM_ID, a.POSTCODE
        FROM PostCode a
        LEFT JOIN SAIL_Link_File1 b ON
        a.POSTCODE=b.PostCode
        GROUP BY a.SYSTEM_ID") # linking to  SNB Donors  (FILE 1) by system ID
View(PostCode_match) # all 111 match
no_match <- sqldf("SELECT SYSTEM_ID FROM SAIL_Link_File1
WHERE SYSTEM_ID NOT IN (SELECT SYSTEM_ID FROM Gender)")
View(no_match) # just checking, but none
```

Figure 5.4: Validating IDEx against the SNB Donors database using sqldf in R studio. Script shows NHS number, Gender, and Postcode validation

IDEx results for the Epi25 cohort

Per-letter scores

In the per-letter scores, the extraction of one annotation containing all required features constituted a positive result.

Variables	Items present	Items extracted	Precision	Recall	F1 score
DoB	761	760	1.00	1.00	1.00
NHS No	657	655	1.00	1.00	1.00
Gender	769	758	1.00	0.99	0.99
Postcode	755	741	1.00	0.98	0.99
Macro Summary			1.00	0.99	1.00

Table 5.3: *Epi25 cohort IDEx validatation against the SNB database – per letter*

Per-person scores

For each category, the correct extraction of at least one annotation containing all required features, per person, was judged as a positive result.

Variables	Items present	Items extracted	Precision	Recall	F1 score
DoB	111	111	1.00	1.00	1.00
NHS No	109	109	1.00	1.00	1.00
Gender	111	111	1.00	1.00	1.00
Postcode	111	111	1.00	1.00	1.00
Macro Summary			1.00	1.00	1.00

Table 5.4: *Epi25 cohort IDEx validation against the SNB database – per person*

Results review

In the per letter scores there was a small number of missed items but this did not have an effect on the overall results, and had no impact on the per person outcomes. Some of the errors are reviewed here.

Date of Birth

In the 771 annotated documents, 761 contained individual's date of birth, with one being missed by IDEx. Seven documents without a date of birth were letters written to patients about their care, where no date of birth was given, which appears to be a usual format for letters sent to individuals about their care. One was a standard clinic letter with a missing date of birth, and two were EEG reports held in the SNB records, where the only reference was the Biobank number. The single false negative was due to a typing error, where date was written in words and a number zero was used instead of 'O'
in October; although not instantly recognizable by a human, it was rejected by the algorithm.

NHS number

NHS number was captured in all but two documents, where it was provided. The two failed annotations were for the documents sourced from the SBUHB document storage system which required some pre-processing to remove formatting before they could be annotated. It appeared that the process was not successful in two cases and the additional tabs prevented the algorithm from extracting the NHS number. The corpus contained many older letters which used patient hospital numbers instead of the NHS number as the reference. Although IDEx is designed to capture hospital numbers for this part of the project, these annotations were not extracted as the validation was carried out against the biobank database which does not store these references. Out of 769 clinic letters (771 documents contained two EEG reports without identifiers), 112 did not have an NHS number.

Gender

IDEx gender capture relies on the use of a reference line at the start of the letter containing a title and/or the first name, and on the use of 3rd person singular personal pronouns within the text (on three consecutive occasions). In letters written directly to patients, rather than to GP or other service providers about the patient, these are missing. For this reason alone, the algorithm failed to capture seven cases. The other four of the 11 missed annotations contained spelling errors in patients' names (no capital letter) and referred to test results without any personal pronouns in the text.

Postcode

Out of 755 Postcodes in the per letter analysis, IDEx failed to capture 14. These included all seven letters written directly to patients, as the patient Postcode algorithm depends on the presence of a clear reference line at the start of the letter. Others were clinic letters without a standard reference line and one letter with a typing error, where letter 'O' was used instead of a zero in the numerical part of the postcode. There were also 16 documents without postcodes, of which two were EEG results notes without any patient information apart from the biobank number, and the others were clinic letters.

5.2.3 Validation of Clinic, Letter, and Record Date extraction

Additional validation of clinic and letter dates extracted from the Epi25 cohort was performed on a random sample of 200 documents. This was the only validation of Letter Date annotation as this entity was added to the pipeline after it was noted that some of the documents in the set did not contain a clinic date. It was also an evaluation of the new Record Date which was created on the bases of clinic and letter dates, specially to deal with such situations i.e. to date the information extracted with the ExECT pipeline when there is no clinic date.

200 letters were randomly selected using Excel RAND function from a table created in r which combined the Clinic and Letter Dates output and created a Record Date for the entire Epi25 cohort document set. These were checked manually in GATE developer, and the results, as True Positives, False Positives, and

False Negatives were noted, and precision, recall and F1 score were calculated in Excel, as shown in Figure 5.5. Within the random sample, 121 letters had Clinic Dates (60.5 %) and 182 had Letter Dates (91.0 %) resulting in Record Date for 198 letters (99.0 %). Missing date annotations (False Negatives) were caused by overlapping phrases i.e. when a Clinic Date and Letter Date merged into one annotation, only one of them was extracted. There were also instances of unusual text formats such as capitalisation e.g. ”CLINIC” which were not contained in the gazetteer, and some spelling mistakes.

Variables	Items present	Items extracted	Precision	Recall	F1 score
Clinic Date	121	118	1.00	0.98	1.00
Letter Date	182	181	1.00	0.99	1.00
Record Date	198	198	1.00	1.00	1.00

Table 5.5: *Validation of Clinic, Letter, and Record dates on a random sample of 200 letters from the Epi25 cohort document-set*

In the full Epi25 set of documents, 287 had no clinic date (37.2 %), 74 (9.6 %) had no letter date, and 3 (1.0%) had neither, meaning that out of 771, 768 (99.6 %) provided a dated record. This was used in the later analysis of the ExECT output for the Epi25 cohort, where Record Date was added to all data tables.

5.3 Chapter summary

- The IDEx pipeline was created to provide a dataset of personal demographic information that can be used for data linkage studies.
- Validation, not including Gender annotations, was performed on the 200-letter set from the original ExECT v1 evaluation.
- An updated version of the pipeline containing algorithms to annotate letter date and patient’s gender was validated against a set of documents for the Epi25 cohort of individuals. This created a dataset which was used as FILE 1 for the SAIL genetic data linkage study.

- The validation of personal demographic information extraction from the Epi25 cohort documents was carried out against the SNB donors dataset, achieving a perfect score for the per person results.

- Record Date was created for the Epi25 cohort dataset based on the extracted Clinic and Letter Date annotations. The extraction of all three dates was evaluated using 200 randomly selected documents. Record date, which is a clinic date or a letter date if the former is missing, should be used in the linkage of ExECT output files.

Chapter 6

EPI25 COHORT PROCESSING - OUTPUT VALIDATION AND ANALYSIS

This chapter describes the outputs from the ExECT pipelines for a large set of clinical documents for a cohort of participants from the Epi25 study. Outputs are processed to create per-person datasets uploaded later to the SAIL databank for data linkage. They are also used for further validation of the pipelines' output against information held in the Swansea Neurology Biobank. An example of analysis of the linked data from the outputs is also given, both for individual participants and for the entire cohort.

6.1 Per person Epi25 cohort dataset

6.1.1 The datasets

Epi25 study

The Epi25 cohort dataset is a set of clinic documents for consented individuals who donated DNA samples to the SNB that were exome sequenced as part of the Epi25 collaboration. The documents were annotated by the NLP pipelines and processed to create detailed datasets to i) supplement clinical data within the SAIL databank for the Genetic linkage study (Chapter 6) ii) further validate the pipeline using the SNB data extract. More detailed description of the Epi25 project is given in section 2.6.1 on page 60.

SNB dataset

The SNB dataset has been derived from two of the SNB's databases, Clinical and Donors. Two separate datasets were produced using an Access query. The Donors dataset provided a set of personal details to validate the IDEx pipeline. The Clinical dataset was used to validate diagnosis, investigations, and febrile seizures and to compare ExECT output for Onset and certain items from patient history. The structure and content of the databases are described in section2.1.3 on page 40.

6.1.2 Documents processing

Clinic letters used were extracted from the SNB records, preprocessed, and converted to plain text files without de-identification. This process is described in detail in Section 2.1.2 on page 39. Documents were saved with an identifier based on the biobank number and a consecutive letter number for each person's set. The final corpus of 771 documents consisted of clinic letters, letters to patients, letters to GP providing investigation results and a small number of file notes containing EEG results.

6.1.3 Pipeline outputs and data table construction

Output files from the NLP pipelines were produced by attaching Groovy scripts at the end of the pipeline for each annotation category in turn. Groovy script sets out all features to be extracted, whereas output parameters such as file format and location are set in the GATE pipeline as other processing resources. 6 separate annotation output files were produced from IDEx (Chapter 4) and 9 from ExECT. All files contained a letter reference column with values derived from the biobank number and the letter number in each person's set. This column was used to create a SYSTEM_ID to allow for linkage by individual. This was carried out using the Microsoft Excel split option by the duplication of the letter reference column and removal of the section indicating letter sequence.

6.2 ExECT output

The ExECT pipeline produces 9 separate annotation files: Diagnosis (epilepsy), Birth History, Onset (epilepsy), When Diagnosed, Epilepsy Cause, Investigations, Prescriptions, Seizure Frequency, and Patient History.

6.2.1 ExECT Epilepsy Diagnosis

1,632 epilepsy diagnosis annotations were extracted from 771 documents for 111 individuals. All annotations contained the full set of features as described in Chapter 3 4.1 on page 83. Following the removal of statements with certainty levels below 4 or 5 i.e., less certain diagnoses (see Materials and Method chapter, section 2.4.5 on page 53; the list of levels is given in appendix B) the number of annotations was reduced to 1585.

Validating the epilepsy diagnosis for each individual was based on the most recent documents within each participant's set. This is because diagnosis may change as epilepsy changes, or become more clarified as more information becomes available with time. For example, a person diagnosed with childhood absence epilepsy at 7 years of age may later have juvenile myoclonic epilepsy, or a diagnosis of focal epilepsy may become more specific, such as temporal lobe epilepsy. The

Diagnosis table was linked, therefore, by letter reference to the Record Date table produced by IDEx, adding a date to all diagnoses, unless the source document was not dated. Record date was used to identify most recent clinic letters.

The extraction of epilepsy type was based on the most recently stated specific epilepsy or syndrome. When these were not given, it was based on the stated seizure type if that could be clearly associated with a type of epilepsy. For example, tonic clonic seizure may occur in generalised or focal epilepsy and may not be used to identify generalised genetic epilepsy, but primary generalised tonic clonic seizure can. Priority was also given to the diagnosis most frequently mentioned within each document. For example, in a clinic letter containing 2 mentions of juvenile myoclonic epilepsy, 1 mention of absence epilepsy, and 5 mentions of absence seizures, juvenile myoclonic epilepsy would be extracted as the diagnosis during the validation, provided the level of certainty was greater than 3.

Epilepsy exclusion criteria Epilepsy diagnoses that were excluded during the validation by epilepsy type were those not uniquely associated with focal or generalised epilepsy and included Epilepsy, Symptomatic epilepsy, and Drug resistant epilepsy.

Seizure exclusion criteria Seizures excluded from the diagnosis validation by epilepsy type were those that can occur in both focal and generalised epilepsy and included Epileptic seizures, Intractable seizures,Tonic clonic seizures, Generalised seizures, and Myoclonic seizures.

The results of the filtration process of diagnosis annotations are shown in Fig.6.1. For 111 participants, 97 had a specific epilepsy diagnosis identified by epilepsy type, nine by seizure type alone, and four had non-specific epilepsy. Non-specific epilepsy diagnoses included 'Epilepsy', 'Tonic clonic seizures', and 'Symptomatic epilepsy'.

The missing epilepsy types are reviewed below. It is important to note that Epi25 submission criteria required clear evidence to support epilepsy type [**?**] and that diagnosis for patients with non-specific epilepsy would have been reviewed

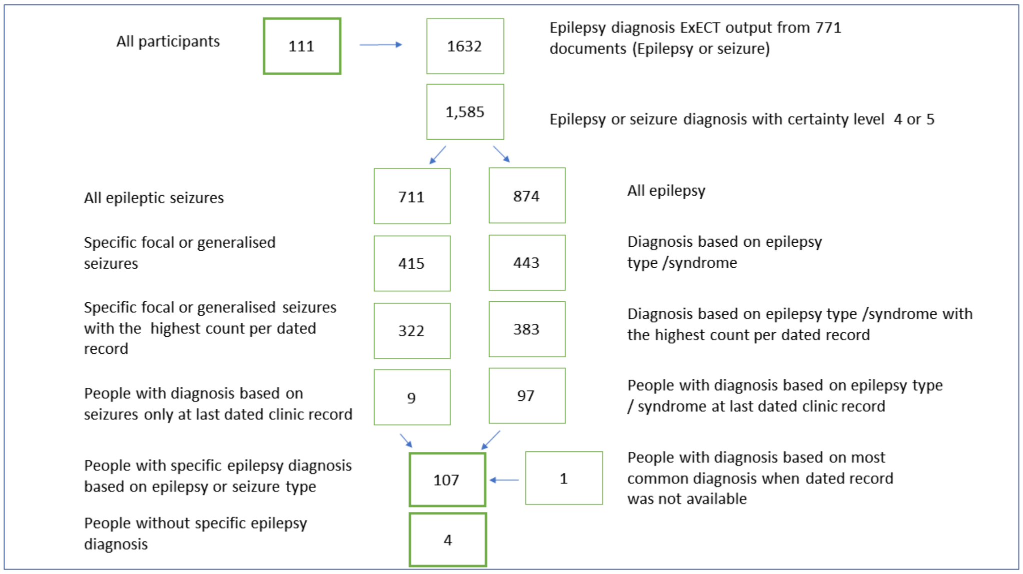

All participants
111
1632
Epilepsy diagnosis ExECT output from 771 documents (Epilepsy or seizure)
1,585
Epilepsy or seizure diagnosis with certainty level 4 or 5
All epileptic seizures
711
874
All epilepsy
Specific focal or generalised seizures
415
443
Diagnosis based on epilepsy type /syndrome
Specific focal or generalised seizures with the highest count per dated record
322
383
Diagnosis based on epilepsy type /syndrome with the highest count per dated record
People with diagnosis based on seizures only at last dated clinic record
9
97
People with diagnosis based on epilepsy type / syndrome at last dated clinic record
People with specific epilepsy diagnosis based on epilepsy or seizure type
107
1
People with diagnosis based on most common diagnosis when dated record was not available
People without specific epilepsy diagnosis
4

Figure 6.1: *Filtering ExECT Epilepsy Diagnosis output by certainty, specific epilepsy type, specific seizure type, most recent dated clinic record, most common diagnosis if dated record is not available.*

and confirmed using clinic letters, EEG, and MRI reports.

Grouping epilepsy CUIs To carry out the validation of ExECT epilepsy diagnosis output against the SNB records, all epilepsy and seizure CUIs used in ExECT gazetteers were grouped by epilepsy type or syndrome. The list contained 71 CUIs for focal and 41 for generalised epilepsy, in addition to specific syndrome CUIs, and a number of CUIs that can be associated with either type of epilepsy. Only CUIs relating to generalised and focal epilepsy were used in the validation of epilepsy type.

Swansea Neurology Biobank epilepsy diagnosis

Information relating to epilepsy diagnosis within the SNB database is held in the Epilepsy Diagnosis and Seizure Type section. This information is not coded, therefore, in order to compare it with the ExECT output all diagnostic terms were assigned an appropriate CUI. This was done manually as the terms used within the database do not match the UMLS PREF terms. For example, focal epilepsy in the database is recorded as ‘Focal’, whereas a corresponding UMLS PREF is ‘Epilepsies, Partial’. Coding was carried out in Microsoft Excel following data extraction. To confirm final epilepsy type for each person’s record, the biobank extract was run against the epilepsy type list based on the ExECT gazetteer, the R script is given in Appendix C.1, fig **??**. This process identified 78 individuals with focal and 33 with generalised epilepsy.

ExECT epilepsy diagnosis validation

ExECT epilepsy diagnosis was validated for each individual against the SNB-held information using precision, recall, and F score (Table 6.1). Epilepsy diagnosis was confirmed for all individuals, focal epilepsy extracted for 72 people and generalised for 35. There were four individuals for whom epilepsy type was not identified by the pipeline and two people with the biobank diagnosis of focal epilepsy which was extracted by ExECT as generalised.

Variable	**Items present**	**Items extracted**	**Precision**	**Recall**	**F1**
All epilepsy	111	111	1.00	1.00	1.00
Focal	78	72	1.00	0.92	0.96
Generalised	33	33	0.94	1.00	0.97

Table 6.1: *Epi25 cohort ExECT output for Epilepsy diagnosis validation against the SNB records.*

Missing epilepsy type A specific epilepsy diagnosis was not extracted by the pipeline for four individuals. Validation against the SNB data shows that all missing diagnoses were for focal epilepsy. Reviewing the ExECT output and the original documents showed that for these individuals there were not enough of the diagnostic details, that ExECT is designed to capture, for ascertaining epilepsy type. For example, focal abnormality on EEG would not be extracted by ExECT which is set to annotate investigation results as normal or abnormal, Table 6.2.

Person	Document type	Diagnostic details
1	3 clinic letters	Unclassified epilepsy, no seizure type
	2 EEG results	Suggestive of focal abnormality
2	2 clinic letters	Generalised tonic clonic seizures
	2 EEG results	normal
3	3 clinic letters	Epilepsy, episodes rather than seizures
	2 EEG results	Suggestive of focal abnormality
4	3 clinic letters	2 mention epilepsy, 1 Temporal lobe epilepsy with certainty level 3 (possible)
	3 EEG results	No diagnosis

Table 6.2: *Epi25 cohort ExECT output, analysis of missing Epilepsy diagnosis.*

Misclassified epilepsy type There were two cases of focal epilepsy in the SNB database which were identified as generalised by ExECT i.e. false negatives for focal and false positives for generalised epilepsy extraction. These are reviewed here.

Case 1 Detailed inspection of the ExECT output shows that the only epilepsy diagnosis given, ‘Idiopathic generalized epilepsy’, was in 2015 and this was the most recent dated record extracted for epilepsy type. Other annotations were all for seizures i.e. ‘Complex Partial Seizure’, ‘Generalized tonic-clonic seizures with focal onset’, and were much more recent. This was confirmed by reviewing the original documents. Had the annotation selection not given priority to Epilepsy Type / syndrome and extracted any epilepsy terms at last clinic date, the output would have been Focal epilepsy.

Case 2 Biobank records for this person were updated in July 2017 with the diagnosis stated as ‘Focal epilepsy’ and seizure type recorded as ‘bilateral convulsive secondary generalised’. ExECT output is dated June 2020 and gives ‘*Idiopathic generalized epilepsy*’, so it appears that in this case the diagnosis has changed. To confirm this, output from six letters available for this individual was inspected. All clinic letters were dated between September 2017 and June 2020, of which three gave the diagnosis of idiopathic generalised epilepsy, which confirms that the diagnosis has been changed from focal to generalised epilepsy.

Alternative processing Another approach to processing diagnosis is to treat epilepsy type/syndrome and seizure type as having equal diagnostic value and select annotations on the basis of the most recent (dated) clinic record. Testing this approach produced tree mismatches, of which one was seen previously (case 2), and two new mismatches (cases 3 and 4 below).

Case 3 This was a diagnosis of focal epilepsy in the SNB records extracted as generalised by ExECT. All other annotations for this individual show '*focal epilepsy*' or 'temporal lobe epilepsy' however, the most recent letter extracted '*absence seizure*' which through CUI was recorded as generalised epilepsy. On reviewing the original clinic letter the phrase was in fact, '*TLE absence seizures*'. This type of seizure is not in the ExECT gazetteer, hence the algorithm annotated a longer match which is absence seizure. TLE absence seizure is not a recognised seizure type and it can be assumed that this was written in error.

Case 4 ExECT extracted focal epilepsy diagnosis whereas the SNB records showed generalised epilepsy. On inspecting the entire output for this individual all other annotations showed '*Idiopathic generalised epilepsy*' and '*absence seizures*'. Reviewing all 12 clinic letters available confirmed that only the last letter referred to focal seizures, without any mention of change in diagnosis. This could be an error or there has been a change, but there does not seem to be enough evidence to support it.

Prioritising diagnosis selection by epilepsy/syndrome seems more appropriate and the results show that it is also more effective in terms of identifying the correct diagnosis as stated in clinic letters. The pipeline can only extract information that is contained in the text.

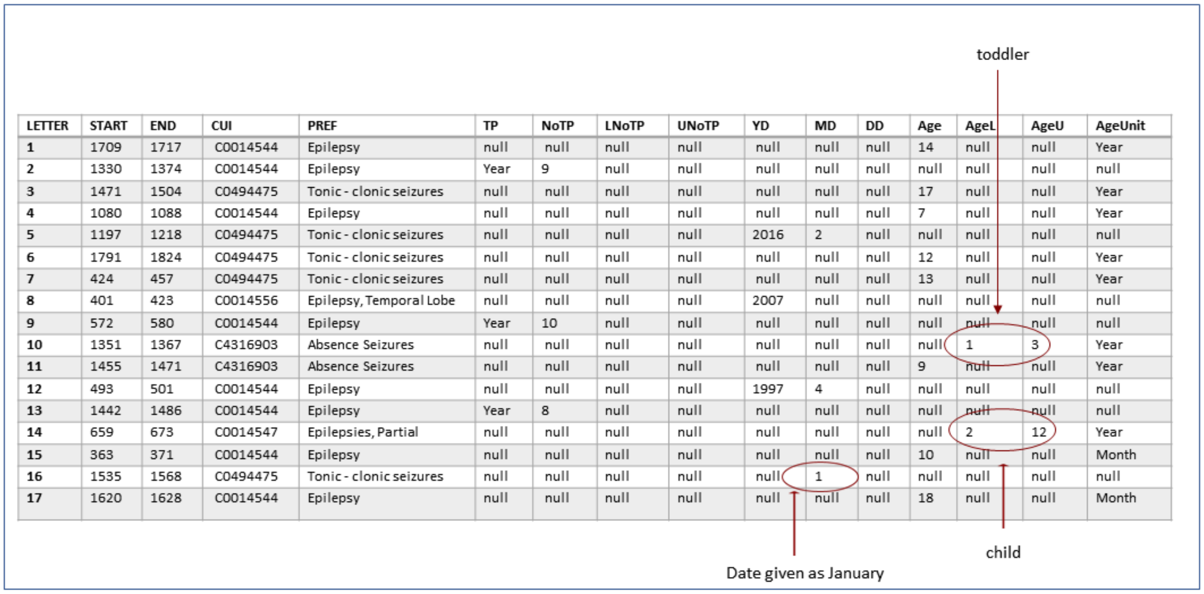
6.2.2 Onset

Unlike diagnosis, seizure frequency, or prescription, epilepsy or specific seizure onset is not mentioned routinely in clinic letters, but rather only when history is reviewed, such as at the initial appointments or when warranted by clinical circumstances. With very few occurrences of onset information, validation on real documents is difficult. The set of 100 letters discussed in Chapter 3 contained only 15 onset references. Having a set of letters for each person increases the chances of onset details being present and captured. From the Epi25 cohort, 43 epilepsy onset annotations were extracted, relating to 30 individuals. As described earlier, onset is recorded as age when epilepsy or specific seizures started, time since they started, or a date The ExECT output reflects this with features containing dates, age, time periods for age expressions i.e. years, months, and time periods for time since onset.

To arrive at a common output from the annotations a degree of processing is required. As age of onset is significant in epilepsy diagnosis and treatment it follows that all annotations had to be converted to age. For this, the output was linked to Date of Birth and Record Date produced by IDEx. Time since onset was subtracted from Record Date to produce Onset Date and the difference between Date of Birth and Onset Date gave age of onset. The full script is provided in Appendix C.2. Fig. 6.2 gives a sample of Onset output from ExECT before precessing and Fig. 6.3 after processing for the same individuals.

Comparing ExECT onset output to SNB records

It is not possible to fully validate ExECT onset output against information held in the biobank. SNB records are based on manual review of participants' notes and on information gained from interviews, where individuals may recall onset of first episodes. It is possible, however, to compare the two sources of information and investigate the differences. Of 111 Epi25 participants, 30 had age of onset extracted by ExECT and 108 had onset recorded in the biobank database, which leaves 78 individuals without onset annotation. Comparing the age of onset in



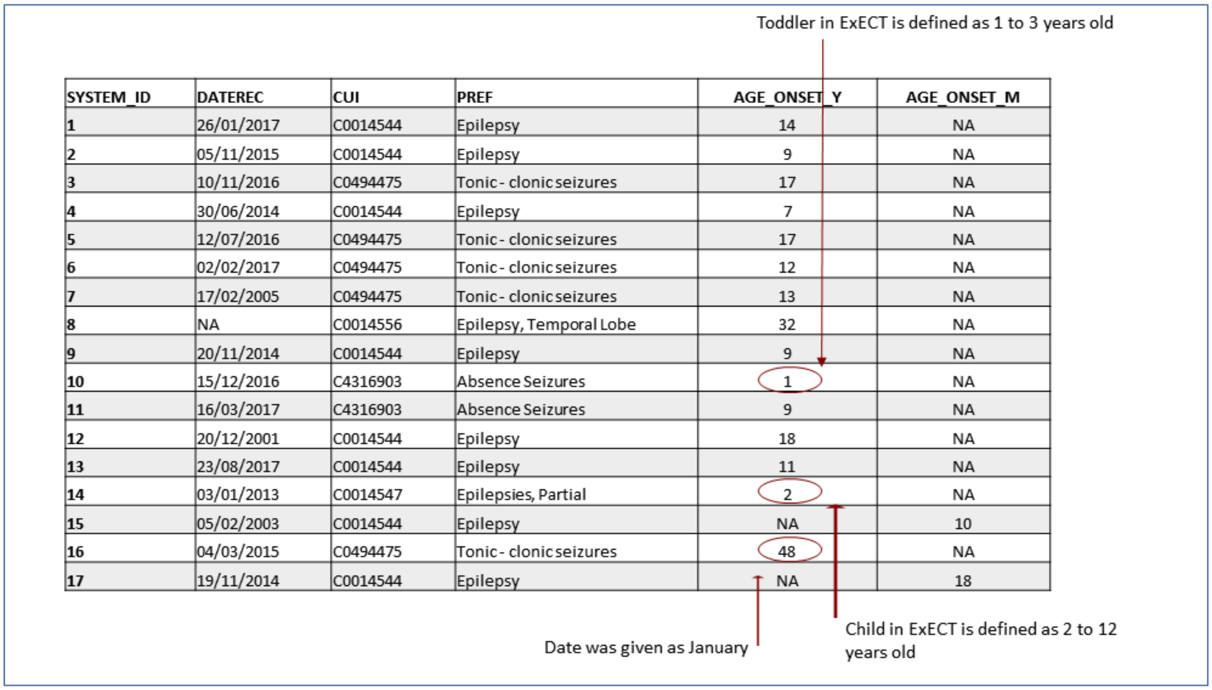
toddler
child
Date given as January

LETTER	START	END	CUI	PREF	TP	NoTP	LNoTP	UNoTP	YD	MD	DD	Age	AgeL	AgeU	AgeUnit
1	1709	1717	C0014544	Epilepsy	null	null	null	null	null	null	null	14	null	null	Year
2	1330	1374	C0014544	Epilepsy	Year	9	null	null	null	null	null	null	null	null	null
3	1471	1504	C0494475	Tonic - clonic seizures	null	null	null	null	null	null	null	17	null	null	Year
4	1080	1088	C0014544	Epilepsy	null	null	null	null	null	null	null	7	null	null	Year
5	1197	1218	C0494475	Tonic - clonic seizures	null	null	null	null	2016	2	null	null	null	null	null
6	1791	1824	C0494475	Tonic - clonic seizures	null	null	null	null	null	null	null	12	null	null	Year
7	424	457	C0494475	Tonic - clonic seizures	null	null	null	null	null	null	null	13	null	null	Year
8	401	423	C0014556	Epilepsy, Temporal Lobe	null	null	null	null	2007	null	null	null	null	null	null
9	572	580	C0014544	Epilepsy	Year	10	null	null	null	null	null	null	null	null	null
10	1351	1367	C4316903	Absence Seizures	null	null	null	null	null	null	null	null	1	3	Year
11	1455	1471	C4316903	Absence Seizures	null	null	null	null	null	null	null	9	null	null	Year
12	493	501	C0014544	Epilepsy	null	null	null	null	1997	4	null	null	null	null	null
13	1442	1486	C0014544	Epilepsy	Year	8	null	null	null	null	null	null	null	null	null
14	659	673	C0014547	Epilepsies, Partial	null	null	null	null	null	null	null	null	2	12	Year
15	363	371	C0014544	Epilepsy	null	null	null	null	null	null	null	10	null	null	Month
16	1535	1568	C0494475	Tonic - clonic seizures	null	null	null	null	null	1	null	null	null	null	null
17	1620	1628	C0014544	Epilepsy	null	null	null	null	null	null	null	18	null	null	Month

Figure 6.2: *ExECT Onset output before processing with letter references substituted with numbers for confidentiality. Annotation features extracted: START, END (start and end of the annotation), CUI, PREF(UMLS concept and phrase), TP (time period), NoTP (Number of time periods), LNoTP (lower number of time periods), UNoTP (upper number of time periods), YD, MD, DD(three date field of year, month, and day date), AgeL (lower age), AgeU (upper age)*

the annotated records against the biobank, 17 were matched fully (age and age unit), eight were close matches i.e. within two years either way, and for five, the reported onset age was different. Annotations for these records were reviewed in more details and from the results that are given in Table 6.3.

It is clear that based on the letter content ExECT has correctly identified onset age for four out of five mismatched records. Even for Person B (Table6.3) who corresponds to SYSTEM_ID 14 in Fig. 6.3, ExECT defines a child as a person aged 2 to 12 and the Onset processing selected the lower age from the range as more significant. This case is similar to the close match for SYSTEM_ID 10, who was described in clinic letter as a toddler i.e. person aged between 1 and 3 years in ExECT terms (with the lower age being selected during post-processing). Person E was the only true error and it appears that the rule used is too sensitive and may have to be revised, but it is interesting to note that a small difference in the



Toddler in ExECT is defined as 1 to 3 years old

SYSTEM_ID	DATEREC	CUI	PREF	AGE_ONSET_Y	AGE_ONSET_M
1	26/01/2017	C0014544	Epilepsy	14	NA
2	05/11/2015	C0014544	Epilepsy	9	NA
3	10/11/2016	C0494475	Tonic - clonic seizures	17	NA
4	30/06/2014	C0014544	Epilepsy	7	NA
5	12/07/2016	C0494475	Tonic - clonic seizures	17	NA
6	02/02/2017	C0494475	Tonic - clonic seizures	12	NA
7	17/02/2005	C0494475	Tonic - clonic seizures	13	NA
8	NA	C0014556	Epilepsy, Temporal Lobe	32	NA
9	20/11/2014	C0014544	Epilepsy	9	NA
10	15/12/2016	C4316903	Absence Seizures	1	NA
11	16/03/2017	C4316903	Absence Seizures	9	NA
12	20/12/2001	C0014544	Epilepsy	18	NA
13	23/08/2017	C0014544	Epilepsy	11	NA
14	03/01/2013	C0014547	Epilepsies, Partial	2	NA
15	05/02/2003	C0014544	Epilepsy	NA	10
16	04/03/2015	C0494475	Tonic - clonic seizures	48	NA
17	19/11/2014	C0014544	Epilepsy	NA	18

Date was given as January

Child in ExECT is defined as 2 to 12 years old

Figure 6.3: *ExECT Onset output after processing with all onset expressions converted to Age in years or months. Letter references have been substituted with numbers for confidentiality.*

tense used (Person D vs Person E) has such a significant effect.

6.2.3 Patient History

Patient History output contains annotations for events or diseases that might have caused or be associated with epilepsy, and co-morbidities which may affect treatment, influence or mimic seizures. The range of features for Patient History annotations is described in Chapter 3 4.2.4 on page 95. 1311 patient history annotations were extracted for the Epi25 cohort, ranging from 1 to 56 per individual, averaging 12 per person. In order to compare these annotations to information in the SNB database only items that are collected routinely for all individuals, such as information on febrile seizures or learning disabilities, can be considered. Others,

Person	Age SNB	Age ExECT	ExECT annotation
A	10 years	10 months	*'X suffers from epilepsy since the age of 10 months.'*
B	1 month	2 years	*'X suffers from intractable focal epilepsy since he was a child.'*
C	4 years	13 years	*'this 23 year old X has been having complex partial seizures for the last 10 years'*
D	14 years	17 years	*'X has had 2 tonic clonic seizures, the first in February 2016'*
E	11 years	48 year	*'X had 2 generalised tonic clonic seizures, the first in January'*

Table 6.3: *Age of onset errors for the Epi25 cohort ExECT output vs the SNB records*

relating to patient history and comorbidities are recorded in additional notes, not to a set standard, but only if they are stated as an additional diagnosis in clinic notes or are reported by patients themselves during the interview. There are no certainty levels used with these statements. Validation of the ExECT output for these annotations is therefore difficult, and same examples are discussed below.

Febrile seizures

History of febrile seizures is recorded in the biobank as present, absent, or not known, with the age of onset and type (whether simple or complex) recorded for confirmed cases, if available. For the Epi25 cohort, there were 34 individuals with known febrile seizure history, of which seven had confirmed presence and 27 had confirmed absence of febrile seizures.

ExECT extracted febrile seizure history for 31 individuals (35 annotations in total). Three UMLS CUIs were used in this process, C0009952 (febrile convulsions), C0751057 (complex febrile seizure), and C0149886 (simple febrile seizure). It should be noted that in clinic letters, unlike diagnosis, seizure frequency, or prescription, history of febrile seizures is not stated routinely, and is only mentioned when patient history is being assessed, such as at the first appointment following referral or when a review is carried out for some clinical reason (continued seizures despite treatment, diagnostic uncertainties, further referral etc.).

Clinic letters for the Epi25 cohort do not contain full sets of documents for each individual (see section 2.1.2 on page 39) and it would not be appropriate to validate the extraction of febrile seizures information based on whether it was captured or not. An alternative approach is to look at the information that has been extracted and validate the outcome i.e. whether the history of febrile seizures was negated or affirmed for the individuals that were present (could be matched) in both datasets. Of the 31 individuals with febrile seizures captured by ExECT only nine had recorded febrile seizure history in the SNB, and it was negated, giving the validation results shown in Table 6.4. These are positive results but the numbers of cases are very small.

Variable	**Items present**	**Items extracted**	**Precision**	**Recall**	**F1**
Negated Febrile Seizure	9	9	1.00	1.00	1.00

Table 6.4: *Epi25 cohort ExECT output for 'No history of febrile seizures' validation against SNB records*

There were no affirmed cases of febrile seizures extracted by ExECT and no comparison with the SNB records could be made.

Comorbidities

As comorbidities are not uniformly recorded in the SNB database the ExECT annotation outputs can only be compared to the biobank rather than validated. The two most commonly recorded comorbidities in the biobank for the study cohort were depression and dissociative seizures; both were recorded for nine out of 111 individuals. ExECT annotations for depression were extracted for 21 individuals, but only three had a biobank record. Dissociative convulsions, also known as non-epileptic seizures, were extracted for 19 individuals with only two having a biobank record.

6.2.4 Investigations

Investigations output from ExECT contains annotations for EEG, MRI, and CT scan results. These are described as normal or abnormal and no specific abnormalities are given. There is no date associated with the investigations as Record Date relates to the clinic letter in which the investigations were mentioned, not when they were performed. For that reason multiple annotations may refer to the same result, as it might be repeated in many letters. Validating investigations may, therefore, be carried out only in terms of a single (per person) result for each type of (normal and abnormal) outcome, and for each type of test. For example, two standard EEGs that are normal will give one normal Standard EEG; one normal EEG and one abnormal EEG would give two results, one normal and one abnormal; and one normal standard EEG and one normal sleep deprived EEG will give two normal results, as from the EEG type it is clear that these were two separate investigations. However, as the numbers of results by EEG type are small (Tables 6.5 and 6.6), it may be more helpful to group the results by outcome only. The same approach was used when comparing the ExECT output to the biobank records for the MRI results. No comparison was made for the CT scans as the number of reports was small, with four abnormal and 11 normal results extracted by the pipeline as compared to 11 and 18 respectively from the SNB.

ExECT EEG output

Out of 670 Investigation annotations extracted by ExECT from the Epi25 cohort document set, 417 referred to EEGs. There were 122 normal and 287 abnormal results,(eight were not stated), with 59 individuals having at least one normal and 98 at least one abnormal result when grouped, disregarding the test type. Results by test type are given in Table 6.5

Result	Standard	Sleep	Prolonged	Ambulatory	VideoTelemetry
Normal	47	5	1	1	5
Abnormal	80	8	1	3	6

Table 6.5: *ExECT output for EEG results from the Epi25 cohort processing, by test type, grouped by individual*

EEG SNB records

The SNB EEG records are slightly different to ExECT annotations of abnormal results, being a little stricter and categorising less certain cases as ‘Other’. This category does not feature in ExECT annotations, but does have an effect when comparing the records between the two sources. Table 6.6 gives the numbers of normal and abnormal results when grouped by individual.

Result	Standard	Sleep	Prolonged	Ambulatory	VideoTelemetry
Normal	37	8	1	1	6
Abnormal	74	20	0	6	11

Table 6.6: *The SNB records of EEG results by test type grouped by individual*

Comparing ExECT EEG result output to the SNB records

The validation of ExECT EEG annotations was carried out on grouped results for standard, sleep deprived, prolonged, ambulatory tests, and video telemetry, per individual for grouped normal and abnormal outcomes, Table 6.9. The grouping is somewhat misleading as the same value is given to a single result as to a number of results of the same outcome per person.

The ExECT’s performance in identifying abnormal results is comparable to that reached in the validation of the overall results (section 4.2.4 on page 92). The results for normal outcomes do not seem that good. This may be caused by

Variable	**SNB**	**ExECT matched (TP)**	**ExECT missed (FN)**	**Precision**	**Recall**	**F1**
Normal EEG	49	24	25	0.50	0.49	0.50
Abnormal EEG	92	73	19	0.87	0.79	0.83

Table 6.7: *ExECT EEG result output for the Epi25 cohort validation against the SNB records*

the selection of letters available, that may contain more up-to-date information for some individuals, but also not contain the full history of investigations for the others.

On reviewing the annotations for the normal results comparison, for the false negative results, no normal result could be found in the letters available i.e. all annotations extracted were correctly annotated as abnormal results by the algorithm, with many letters not containing any references to EEG reports. For the 24 false positives (not shown in the table), 12 were in fact correct extractions and 12 were true errors. The main reasons for the errors was the reporting of two investigations with different results in a single phrase, and the inclusion of long extracts of investigation reports which referred to parts of the recording as normal, although the overall result was not.

Inspecting the abnormal results, among the 19 false negatives, for 14 individuals there was no mention of an EEG result in the documents available, three had a normal EEG result reported and correctly extracted, and two were errors, one caused by wrongly annotating an abnormal result as normal, the other, by a reference to EEG as ‘brain recording’ which is not included in the ExECT gazetteer, and was missed.

For the 11 false positives (not shown in the table), ten were identified as correct annotations of abnormal EEG outcomes, and only one was annotated in error.

Comparing ExECT MRI result output to the SNB records

There were 299 MRI results extracted by ExECT, of which 155 were normal, 73 were abnormal, and one was ‘unknown’. The SNB database provided 128 MRI results, 80 normal, 39 abnormal, three uncertain, with six without a stated outcome. The MRI results extracted by ExECT were compared to the SNB database records for grouped reports of the same outputs, normal/abnormal per individual, as shown in Table 6.8.

Result	**SNB**	**ExECT**
Normal	73(80)	64(155)
Abnormal	34(39)	28 (73)

Table 6.8: *MRI results extracted by ExECT and those recorded in the SNB database grouped by outcome per individual with the figures in brackets referring to the overall number of reports for each outcome group*

As the results extracted by ExECT are not dated, validation may be performed only on grouped results, where a single mention may be compared to many mentions for the same outcome, as described in the opening part of this section, on page 132. The results for the MRI extraction validation against the SNB records for each outcome are shown in Table 6.9.

Variable	**SNB**	**ExECT matched (TP)**	**ExECT missed (FN)**	**Precision**	**Recall**	**F1**
Normal MRI	73	53	20	0.82	0.57	0.67
Abnormal MRI	34	15	19	0.54	0.44	0.48

Table 6.9: *ExECT MRI result validation against the SNB records, per individual grouped results for normal and abnormal outcome.*

These figure did not match the performance shown during the validation, Fig. **??** on page **??**, and for this reason they required further investigating.

For normal MRI results, out of the 20 false negatives, 14 could in fact be considered as correct results, since there was no evidence in the letters available for normal MRI annotations to be made. There were six true false negatives, where the normal results were missed; some were caused by an unusual way of reporting the result, for example the MRI was referred to as '*MR brain*' (not in the gazetteer) or the unusual way of reporting the result, for example '*MRI – no evidence of MDS*'.
For the 11 false positive results, seven were, on inspection, correctly annotated normal results, and four were true false positives. Some of these were caused by the investigation results for EEG and MRI being given as one phrase, as in '*Investigations: EEG – normal MRI –- Old right occipital ischaemic lesion.*', with ExECT picking up '*normal MRI*' as an outcome, and by the ambiguity of the result itself, as in '*MRI – Mild generalised cerebral and cerebellar volume loss, no specific abnormality*', which was annotated as normal, but may be considered as abnormal.

Among the abnormal results, in the 19 false negatives, 11 were identified as correct i.e. there were no abnormal MRI results in the letters available. For some individuals the number of letters was small and although from the diagnosis, it was evident that at some stage an MRI scan was performed, it was not referred to in the documents available. There was also eight true false negatives, where ExECT missed an abnormal result. Some of these were due to the type of abnormality reported, which was not present in the results' gazetteer, but also in the way the results were given, as in the example quoted above.
The review of the false positive cases for abnormal results (13), identified six correct annotations made by ExECT of the abnormalities reported, and seven true false positives. For the correct annotations, the letters were quite recent and referred to the reports that were more recent than the records in the biobank.
The reason for the true errors is mainly due to the hypothetical statements being annotated as results, and where the diagnosis of epilepsy is followed by an MRI

result with no punctuation.

6.3 Linking ExECT outputs for longitudinal analysis

This section explores the output for two annotation types, Seizure Frequency and Prescription. The detailed description of the output structure and features is given in Chapter 3, here it is explored further, as a resource for longitudinal analysis using data extracted from the Epi25 cohort documents.

6.3.1 Seizure Frequency

Achieving seizure freedom without significant side effect is the ultimate aim of epilepsy treatment. Seizure monitoring is therefore an integral part of patient consultation. Any changes in seizure frequency, severity, or aetiology are recorded in clinic notes. There is no prescribed approach to this recording as it is based on patients reporting their seizure experience. Apart from the most categorical statements such as ‘seizure free’ or ‘not seizure free’ reported seizure frequency is difficult to measure and compare between different individuals or from one clinic appointment to another for the same individual.

Frequency may be recorded as a number of seizures per specific period of time e.g. 3 per week, or a number of seizures during a distinct period e.g. 6 last month, or since a particular point in time e.g. 10 since June, and so on, as discussed in section 4.2.4 on page 102. The Seizure Frequency output from ExECT reflects this complexity, and further processing is needed to translate the output produced into a measurement of seizure frequency that can be used when comparing seizure activity between different individuals or the same person over time.

ExECT output

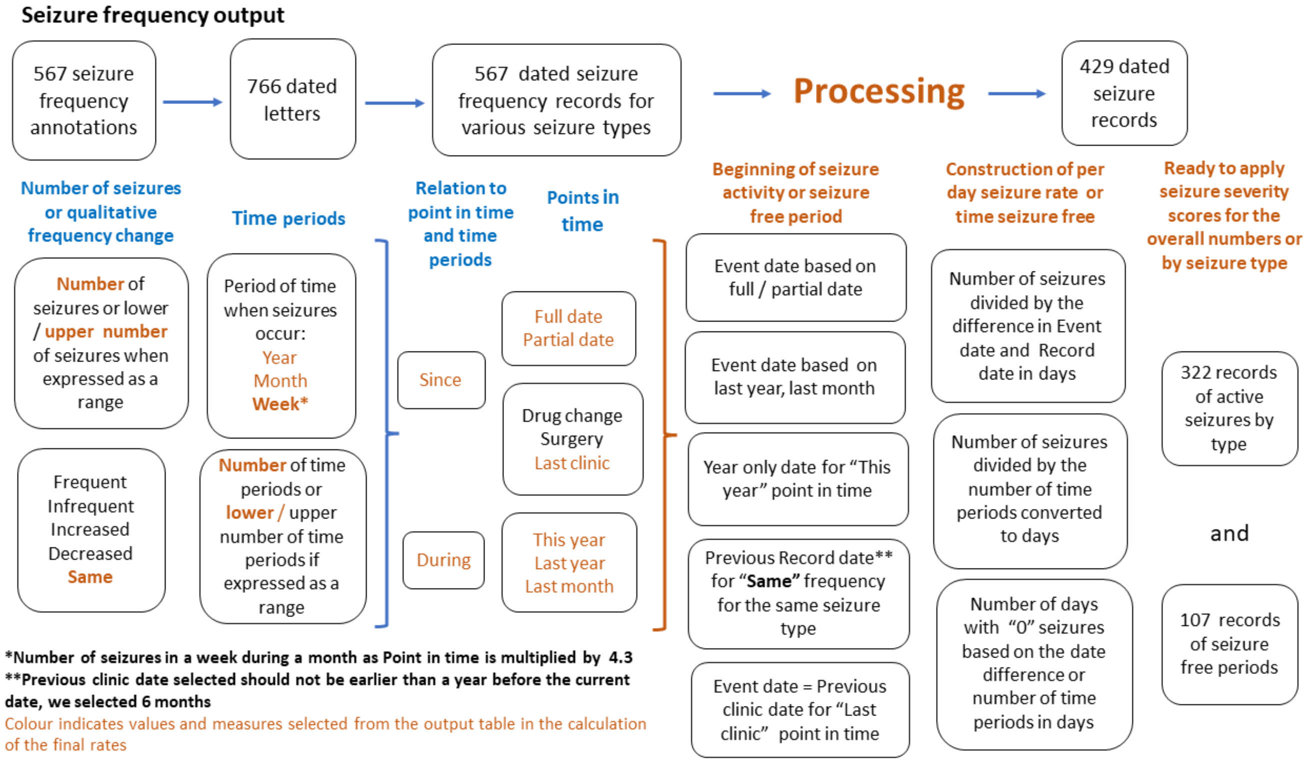
For the Epi25 cohort there were 567 dated annotations of seizure frequency. The dated records were created by linking the ExECT Seizure Frequency output with

the Record Date table produced by IDEx processing (section 5.5, on page 117).
For each seizure type (its CUI), the number of seizures per time period was calculated for the stated frequency. This was converted to seizures per day after all time periods were converted to days. For time periods expressed as a range e.g. per 10–20 days, the lower number was selected as the denominator, and when seizures were expressed as a range, the higher number was selected as the numerator, so as to never underestimate the reported rate.
For all references to time (since date) time periods were calculated and used in frequency per time. For statements giving '*last clinic*' and '*same*' (presumably referring to the last clinic) it was assumed that the time since the last clinic was not longer than 6 months and a reference date was constructed to calculate seizure frequency. For this analysis, the only seizure frequency excluded was that reported as 'decreased', 'increased', 'frequent', or 'infrequent'.
The final output was a table of dated seizure frequency, by seizure type, and of seizure free periods, which can be compared between patients and over time. This process and the resulting numbers of dated records is illustrated in Fig. 6.4, and the step by step R script used to perform the processing is given in Appendix C.3.

This transformation of different seizure frequency expressions may be converted to scores. In this case, seizure severity assessment scores developed by Fitzgerald et.al. [215] were used, which range from 1 (seizure free for more than 2 years) to 7 (more than one seizure per day). The Seizure Severity Scores (SSS) were reworked for daily rate equivalents, Fig. 6.5, and applied to the daily rate output from the seizure frequency processing.

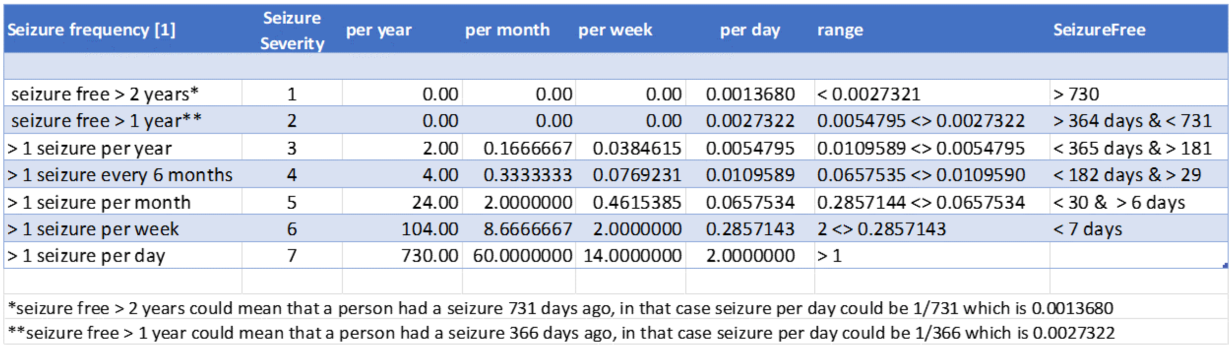
This produced 429 dated SSS by seizure type, allowing for observation of changing seizure pattern for single individuals over time as illustrated in fig 6.6, showing seizure history for a single individual from the Epi25 cohort, Person One.

At the beginning of the period (first clinic records available for the analysis) they experienced generalised tonic-clonic seizures and some non-specific seizures, and although these were reduced, with a 2-year period of seizure freedom reported in 2017 (severity score = 1), seizures returned, with tonic-clonic convulsion being reported at the same clinic visit. From inspecting the output itself it appears


Seizure frequency output
567 seizure frequency annotations
766 dated letters
567 dated seizure frequency records for various seizure types
Processing
429 dated seizure records
Number of seizures or qualitative frequency change
Time periods
Relation to point in time and time periods
Points in time
Beginning of seizure activity or seizure free period
Construction of per day seizure rate or time seizure free
Ready to apply seizure severity scores for the overall numbers or by seizure type
Number of seizures or lower / upper number of seizures when expressed as a range
Period of time when seizures occur: Year Month Week*
Since
Full date Partial date
Drug change Surgery Last clinic
Event date based on full / partial date
Event date based on last year, last month
Year only date for "This year" point in time
Previous Record date** for "Same" frequency for the same seizure type
Event date = Previous clinic date for "Last clinic" point in time
Number of seizures divided by the difference in Event date and Record date in days
Number of seizures divided by the number of time periods converted to days
Number of days with "0" seizures based on the date difference or number of time periods in days
322 records of active seizures by type
and
107 records of seizure free periods
Frequent Infrequent Increased Decreased Same
Number of time periods or lower / upper number of time periods if expressed as a range
During
This year Last year Last month

*Number of seizures in a week during a month as Point in time is multiplied by 4.3
**Previous clinic date selected should not be earlier than a year before the current date, we selected 6 months
Colour indicates values and measures selected from the output table in the calculation of the final rates

Figure 6.4: *Processing of the ExECT Seizure Frequency output from the document corpus for the Epi25 cohort. Various sections of the ExECT output are described in blue, whereas the values used in the construction of the additional dates and rates and the comments on this process are shown in orange*



Seizure frequency [1]	Seizure Severity	per year	per month	per week	per day	range	SeizureFree
seizure free > 2 years*	1	0.00	0.00	0.00	0.0013680	< 0.0027321	> 730
seizure free > 1 year**	2	0.00	0.00	0.00	0.0027322	0.0054795 <> 0.0027322	> 364 days & < 731
> 1 seizure per year	3	2.00	0.1666667	0.0384615	0.0054795	0.0109589 <> 0.0054795	< 365 days & > 181
> 1 seizure every 6 months	4	4.00	0.3333333	0.0769231	0.0109589	0.0657535 <> 0.0109590	< 182 days & > 29
> 1 seizure per month	5	24.00	2.0000000	0.4615385	0.0657534	0.2857144 <> 0.0657534	< 30 & > 6 days
> 1 seizure per week	6	104.00	8.6666667	2.0000000	0.2857143	2 <> 0.2857143	< 7 days
> 1 seizure per day	7	730.00	60.0000000	14.0000000	2.0000000	> 1	

*seizure free > 2 years could mean that a person had a seizure 731 days ago, in that case seizure per day could be 1/731 which is 0.0013680
**seizure free > 1 year could mean that a person had a seizure 366 days ago, in that case seizure per day could be 1/366 which is 0.0027322

Figure 6.5: *Severity Severity Score calculation for seizure frequency expressed as rate per day*

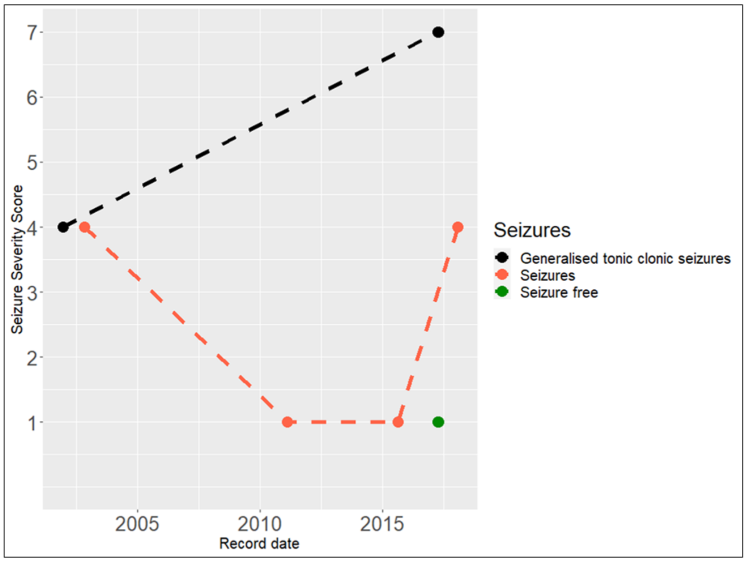


Figure 6.6: *Severity score by seizure type and record date over 11-year period, Person One. The dashed line indicates the direction, with the points representing seizure events.*

that following a period of prolonged seizure freedom Person One experienced a cluster of 5 generalised tonic clonic seizures in one day and some less frequent non-specific seizures in the ensuing period. Looking at the clinic letter from April 2017 confirms the accuracy of ExECT output, it reads: *'X had been seizure free for 2 years until February of this year. Then out of the blue X had a build up of jerks which culminated in X suffering 5 generalised tonic clonic seizures over the course of a day'*. And a letter from February 2018 stating *'X has had 2 further seizures in December and January. Both were clear generalised tonic clonic seizures preceded by a build up of jerks.'*, which is reflected in the increased score for non-specific seizures, as ExECT did not capture the seizure type clarification in the second sentence.

The line on the plot represents a general direction and does not reflect continuous seizure presence, as the points indicate the frequency rate reported at a point in time. This is just one way the output can be interpreted, for example one could use a rate per time period without implementing the severity scores, but this would make it more difficult to compare the periods of seizure freedom and active seizures together.

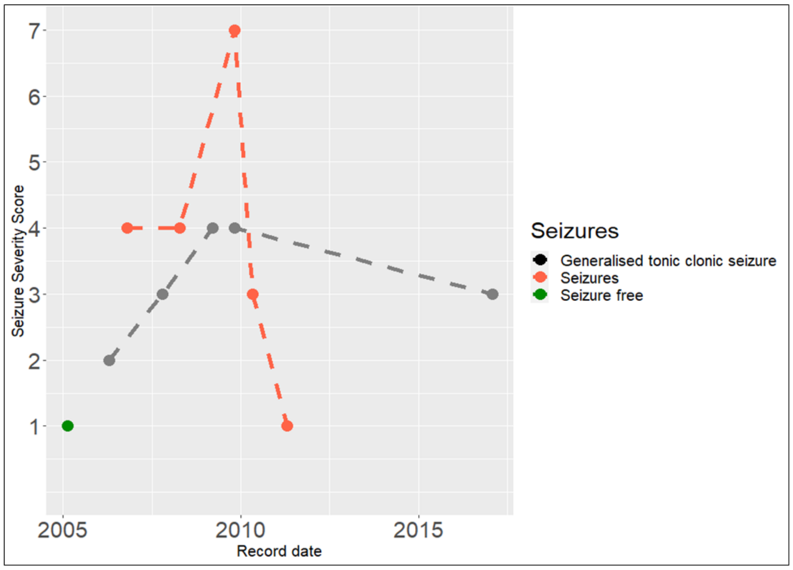
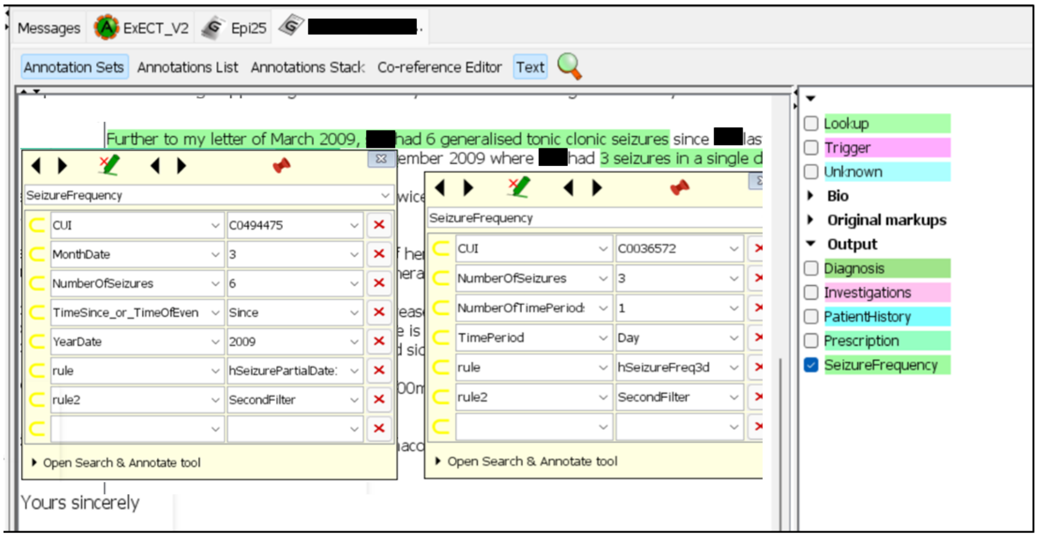


Figure 6.7: *Severity score by seizure type and record date over 13-year period, Person Two. The dashed line indicates the direction, with the points representing seizure events.*

Another example is Person Two, Fig.6.7, who at the beginning of the period, in 2005 (the earliest available document) was seizure free but then reported generalised tonic-clonic seizures and soon after that non-specific seizures. There are no seizure frequency records for this individual between 2011 and 2018, at which point generalised convulsion were reported.

From the documents available the sequence of events can clearly be followed, in 2005 '*X presented with generalised tonic clonic seizures at the age of 13 years....X has been seizure free for the last 7 years. In August 2005 X decided to try to come off anti-convulsants. However, in October was admitted with a generalised seizure'*. In 2006 clinic *'X had further episodes with a generalised tonic clonic seizure in October and has had five generalised tonic clonic events in total the last being in February 2006'*. In October 2009 it is noted *'X had 6 generalised tonic clonic seizures since last visit. The last episode was on 23rd September 2009 where X had 3 seizures in a single day'*. ExECT annotation in GATE developer for this clinic is shown in Fig. 6.8, with the left-hand side annotation box giving features of the first seizure frequency highlighted in green (generalised tonic clonic seizures) and the right-hand side annotation box detailing features of the second annotation (seizures).



Further to my letter of March 2009, [redacted] had 6 generalised tonic clonic seizures since [redacted] las

ember 2009 where [redacted] had 3 seizures in a single d

Yours sincerely

SeizureFrequency

CUI	C0494475
MonthDate	3
NumberOfSeizures	6
TimeSince_or_TimeOfEven	Since
YearDate	2009
rule	hSeizurePartialDate:
rule2	SecondFilter

SeizureFrequency

CUI	C0036572
NumberOfSeizures	3
NumberOfTimePeriod:	1
TimePeriod	Day
rule	hSeizureFreq3d
rule2	SecondFilter

Figure 6.8: *ExECT V2 seizure frequency annotation for Person Two in GATE developer, C0494475 = Tonic-clonic seizures, C0036572 = Seizures*

In the following clinic letters it is reported *'X had a single seizure since the last visit at the end of March 2010.'* and then a year later *'Further to my letter of April 2010, X had a single seizure in the last one year'*. The next and the last available document was from 2017 and it stated *'X's epilepsy is reasonably well controlled on the above medication. X is having 2 to 3 generalised seizures in a year.'*; this is clearly captured by the severity Score of 3 (more than 1 but less than 4 seizure per year) for that period.

The last example of seizure frequency output recorded over a period of time presented as seizure severity scores for specific seizure types is given for Person Three Fig.6.9

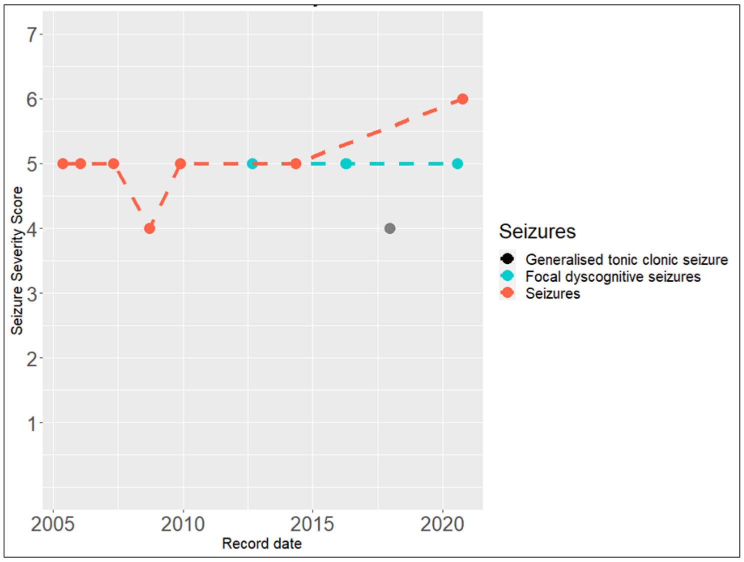


Figure 6.9: *Seizure Severity Score by seizure type and record date over 15-year period, Person Three. The dashed line indicates the direction, with the points representing seizure events.*

The processed seizure frequency output was compared with the ExECT annotations and the original clinic letters. For the first 7 years of epilepsy records available for this individual seizure type was not given when frequency of episodes was discussed, although it was mentioned in the diagnosis. Frequent seizures (more than 1 per month) were reported, including nocturnal episodes, in 2006. A small reduction, to one a month, was reported in 2008. In 2012 a specific seizure type was mentioned for the first time *'X is having approximately 8 nocturnal complex partial seizures per month.'* from which ExECT extracted focal dyscognitive as it is more specific than nocturnal seizures. Clinic letters continue to report high frequency; in 2014 *'X was having approximately 2 seizures per week.'* and in 2016 *'Now X is having approximately 4 complex partial seizures per month.'*

In 2017 it was reported that *'X continues to have 4 to 5 complex partial seizures with secondary generalised tonic clonic seizures per month.'* which was extracted as focal to bilateral convulsive seizures with severity score of 4, as ExECT did not see the range of 4–5 as applying to the second type of seizures, annotating them as occurring once a month. In 2020 *'X continues to have approximately 4 complex*

partial seizures per month. No generalised tonic clonic seizures.', from which focal dyscognitive seizures were extracted (and a severity score of 5 calculated) but not GTCS as these were negated and no time period was given for the seizure-free period.
The last seizure frequency extracted for non-specific seizures is an error as it was derived from *'For the past 10 years X's seizures have been exclusively at night'* and annotated as one per day, which is clearly not correct. Precision of seizure frequency annotations when tested on the gold standard letter set was 78% for all features extracted. In this example, although the fact of having seizures is correct, the features are not. It is often the case when the rules are being developed that not all possible phrases are tested and in some cases rules produce an unexpected outcome. In general it would seem that ExECT correctly identifies seizure type and trends in seizure frequency.

The three individuals reviewed were selected on the grounds of the longest dated treatment record and the largest number of dated documents with extracted seizure frequency. Some of the records were excluded from the analysis as seizure frequency was recorded only as an increase/decrease or frequent/infrequent. It would be possible to incorporate these in the analysis but impossible to compare with frequencies expressed as numbers of events per specific time period. Severity scores used make the comparison between active seizures and seizure free periods possible for different seizure types which gives a good indication of the epilepsy severity.

ASM and linkage with seizure frequency

As for seizure frequency, the Prescriptions output from ExECT was linked to the Record Date file created using IDEx. There were 1,206 dated annotations of prescriptions extracted for the Epi25 cohort. Annotations consist of LETTER (document ID), START (start of annotation span), END (end of annotation span), CUI (UMLS CUI), NAME (Drug name , generic or brand), DOSE (quantity), UNIT (unit of measurement), FREQUENCY (how often the given dose is taken), SYSTEM_ID (person identifier, created from the letter reference), DATEREC

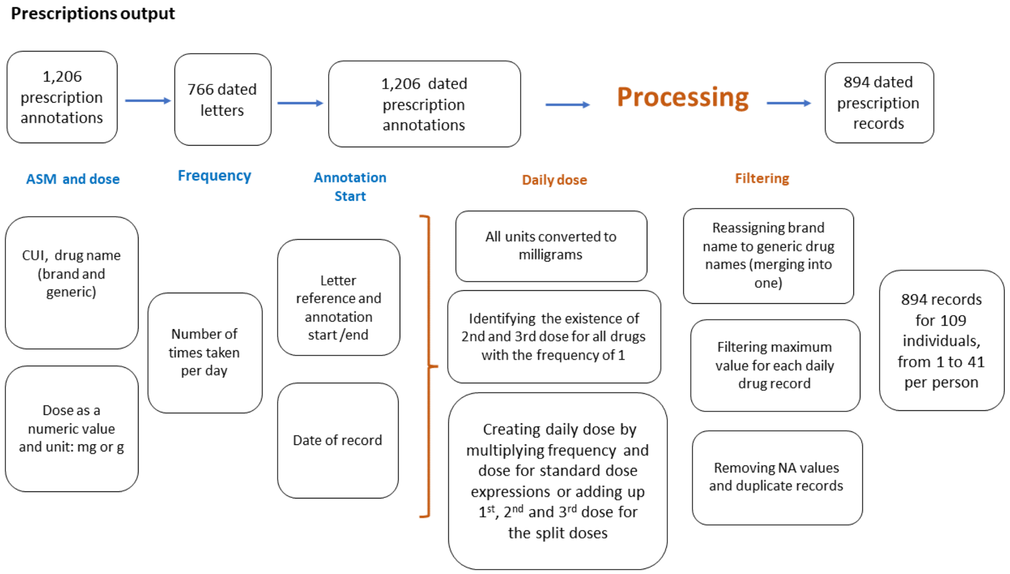

Prescriptions output
1,206 prescription annotations
766 dated letters
1,206 dated prescription annotations
Processing
894 dated prescription records
ASM and dose
Frequency
Annotation Start
Daily dose
Filtering
CUI, drug name (brand and generic)
Dose as a numeric value and unit: mg or g
Number of times taken per day
Letter reference and annotation start /end
Date of record
All units converted to milligrams
Identifying the existence of 2nd and 3rd dose for all drugs with the frequency of 1
Creating daily dose by multiplying frequency and dose for standard dose expressions or adding up 1st, 2nd and 3rd dose for the split doses
Reassigning brand name to generic drug names (merging into one)
Filtering maximum value for each daily drug record
Removing NA values and duplicate records
894 records for 109 individuals, from 1 to 41 per person

Figure 6.10: *Precessing of the ExECT prescription output from the document corpus for the Epi25 cohort. Sections of the ExECT output are listed in blue, whereas the processing elements have orange headings. 'NA'* are NULL values in R.

(a record date created by linking to the IDEx output). Since drug doses can be presented in different formats, for example, 'carbamazepine 600mg bd', as one annotation or 'lamotrigine 200mg mane and 300mg nocte' as two separate annotations, different calculations are needed during the post-processing stage to arrive at a total daily dose for each ASM, as shown in Fig. 6.10, and the R script used in processing of prescriptions is given in Appendix C

Having converted all units of measurement to mg, the full daily prescription for multiple doses of the same quantity of drug was calculated by multiplication (DrugDose × Frequency). Different quantities of medication given in one day were added up after checking that they referred to the same mention of the ASM within the letter, i.e. using annotation 'START'. There were instances where only one part of the daily dose was extracted and that record was not included in the final analysis as this could have suggested a dose reduction. As current prescription can be mentioned more than once within a clinic letter, in the final stage of processing, a maximum daily dose for each ASM mentioned was extracted, giving 894 dated records for 109 individuals, ranging from 1 to 41 per person. From this output it is possible to investigate individuals' treatment history with details of ASM changes, increasing and decreasing doses over a long period of time. As an example the prescription history of the three individuals reviewed previously for seizure frequency can be investigated in more detail.

To be able to compare different ASM, the prescribed dose may be presented as a proportion of the recommended daily dose, and here the daily dose as recommended by British National Formulary (BNF) [216] was used for each drug. The ASM rates were then plotted along a normalised Seizure Severity Score so that they could be compared on a common scale.

For Person One there are 12 dated prescription records from 2001 to 2018 with a long gap between 2002 and 2011 and a four-year gap from then on until 2015. On reviewing ExECT output and individual letters it is evident that until 2018 they were treated exclusively with lamotrigine (sometimes prescribed as Lamictal). Being seizure free, Person One was discharged to GP care in 2011 but was re-referred with an increased number of seizures around 2017 and a suggestion was made to start sodium valproate. This was the only record for VPA in the documents available and since it was not given as a current medication it was not annotated by the pipeline. In 2018 levetiracetam was added. This timeline is shown in Fig. 6.11. A direct link between the ASM prescription and the increased number of seizures is evident from the plot and is clearly confirmed by the description of events in the clinic letters reviewed.

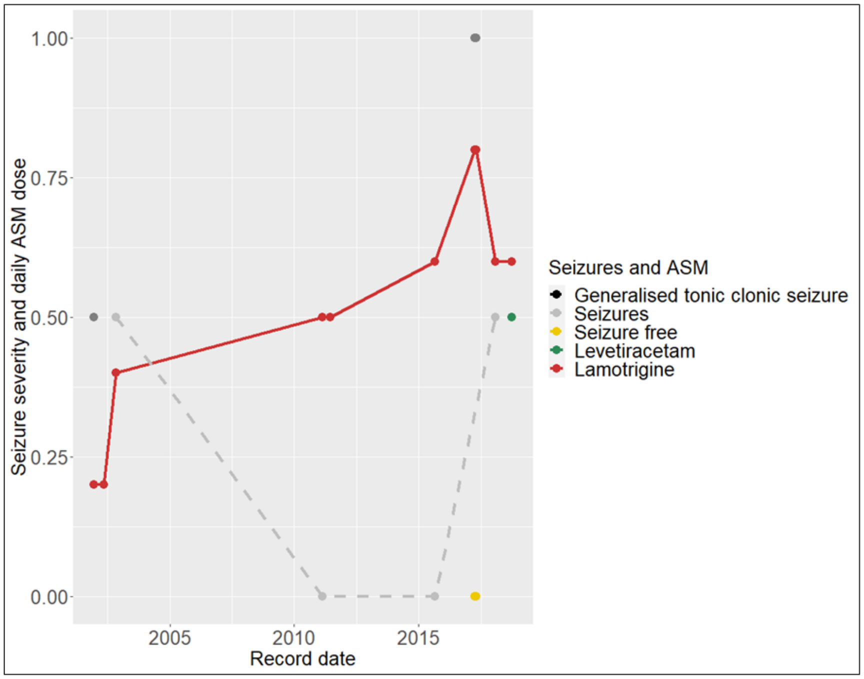


Figure 6.11: *ASM as a proportion of the daily recommended dose and a normalised SSS for Person One.*

Person Two had 26 dated prescription records from 2005 to 2017 with a 6-year interval between 2011 and 2017. From the annotations and the original clinic letters, it appears that they tried four different ASMs, and from the first records available it looks that the document set starts when they were just reducing Primidone (as Mysoline). At the same time they were prescribed Gabapentin and soon after that lamotrigine (also recorded as Lamictal). Gabapentin was stopped in 2007 and around the same time levetiracetam was commenced. Person Two was still taking lamotrigine and levetiracetam in 2017 when our records stopped. Full ASM history, shown as a proportion of the recommended daily dose for each drug, is plotted together with the normalised SSS in Fig. 6.12.

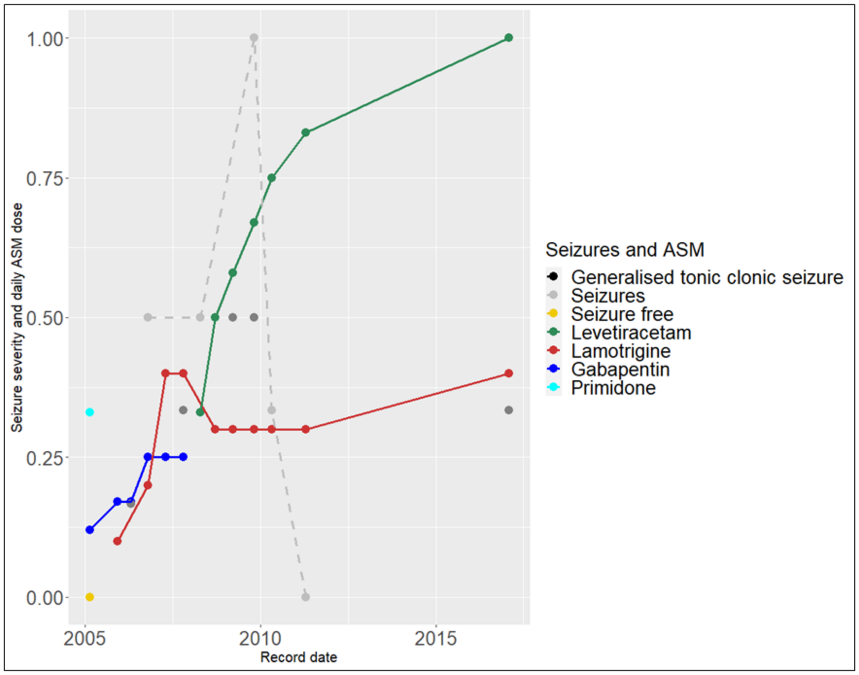


Figure 6.12: *ASM as a proportion of the daily recommended dose and a normalised SSS for Person Two.*

This review identified two errors in the prescription record extraction. Both of them relate to the earlier documents when a more structured practice of stating individuals' medication, in a separate paragraph and/or a list, with a subheading just after diagnosis, was less common.

In a 2005 letter, ExECT treated current medication as a past tense statement (past medication) *'Further to my letter of February 2005, X has been very cautious in tapering off the anti-convulsants and is still taking Mysolin 125 mg twice a day.'* and as a future instruction *'X will cut down Mysolin by 125 mg so that X takes Mysolin 125 mg once at night for the next 1 month and then stop it completely.'*. As past prescriptions are not annotated ExECT misses the first annotation. The second was automatically given a hypothetical status by the context algorithm and as such not annotated. The first example is a definite error on the part of ExECT, the second not totally as this statement is slightly ambiguous; it does not clearly state the current level of Mysolin taken by X.

The second error was again in one of the older letters, from 2006, which states *'As a result of this X had to be commenced on lamotrigine of which the current dose is 50 mg in the morning and 25 mg at night time.'* which was considered a future statement by the context algorithm because of the word 'commenced', but should not have been, as the current ASM with the dose is clearly stated, although in a way that may be difficult to annotate.

Apart from these errors, prescription details extracted by ExECT gave a very accurate record of the ASM history for Person Two and in keeping with the results from the validation with F1 score of 0.93, discussed in section 34.2.2 on page 87. The ASM records reflects that of the SSS plotted alongside, where the increase in the ASM dose corresponds with the increased seizure activity from 2017.

The final example is Person Three, with 41 dated prescription annotations extracted, 5 ASMs from 2004 to 2020, with a 2-year interval between 2017 and 2020. Reviewing the letters and the corresponding annotations has identified a number of missing prescriptions. Three of these relate to a misspelling of carbamazepine, as in *'Medication: Carbarmazepine Retard 400 mg twice a day'* and are all from 2004. As ExECT is based on gazetteers for term identification it is intolerant of different / incorrect spelling. Introduction of common spelling variations may solve the problem although it is difficult to predict all possible errors and there may be a risk of the final word being reassigned to another term. This is something that is being explored for future versions of the pipeline.

Another failed annotation was brought about by the missing (but implied) unit of measurement following the second ASM dose of the day *'Keppra 1000 mg in the morning and 1250 in the evening'*. ExECT rules for annotating prescriptions rely on the presence of a unit of measurement to identify the correct dose. This is important as medication may be given in grams, milligrams, or millilitres. In this case, it is unlikely that the second dose of the day had been prescribed with a different unit of measurement and it might be that a rule is needed for such cases. However, there are examples when mixed units are used *'Keppra 1 g in the morning and 750 mg in the evening'* so any rules would have to be very carefully tested.

The incorrectly missed prescription in 2010 was probably triggered by 'started' which is one of the prescription context words implying new treatment. Here it affected the later statement, of the type that is usually correctly extracted, *'They have started tapering off Keppra and at present X is taking Keppra 1 gram a day'*. In general, apart from the above errors, ExECT performed very well in extracting ASM history for Person Three. When linked with the SSS, and bearing in mind the error relating to the last reported frequency (non-specific seizures), a picture of difficult to control seizures seems clearly visible, (Fig. 6.13) It has to be noted that there were three records that could not be included in the analysis as clinic/letter dates were missing, and that the seizure frequency only includes quantifiable expressions, so a number of reports might have been excluded from this analysis.

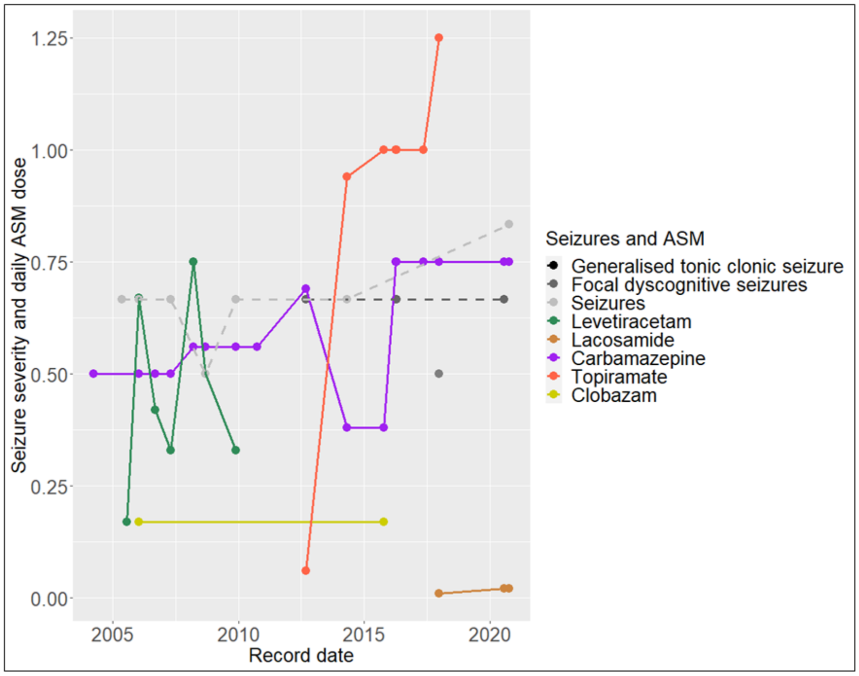


Figure 6.13: *ASM as a proportion of the daily recommended dose and a normalised SSS for Person Three.*

6.3.2 Linking seizure frequency and ASM outputs for the entire cohort

This section explores how the prescription and seizure frequency NLP outputs could be linked and investigated for possible patterns in population-based analysis, using the entire Epi25 cohort as an example.

In order to present each dated record of linked medication and seizure frequency, the number of ASMs taken was mapped to the SSS, first for monotherapy and polytherapy, and then by the ASM type. Fig. 6.14 shows the mapping of monotherapy and polytherapy on the SSS for all seizure types. There were 122 dated SSS and ASM records for monotherapy and 207 records for polytherapy. The data do not show the rate of the dose given; this was not possible for polytherapy, whereas for monotherapy, there were only five records (four individuals) on the maximum recommended daily dose.
It is difficult to see any clear pattern in the data, apart form, that a small number of records with the highest SSS (daily seizures) were seen for monotherapy, but individuals on polytherapy still experience frequent seizures, more than 1 per month.

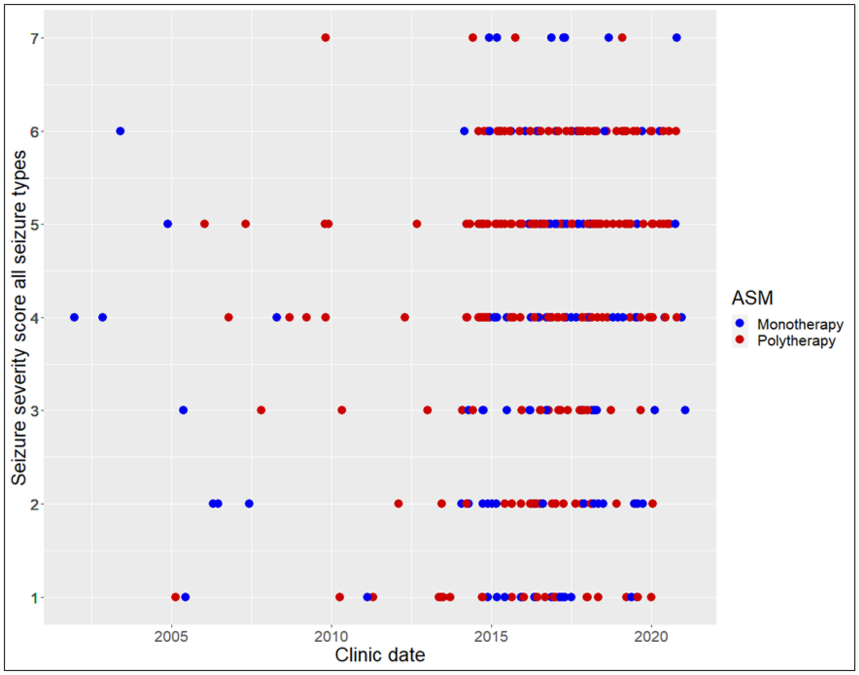


Figure 6.14: *ASM monotherapy and polytherapy and seizure severity score for all seizures types.*

The pattern seen is not dissimilar to that observed when GTCS are selected, as shown in Fig. 6.15, although the highest frequency is observed for the two records of monotherapy. There were 87 dated linked records, with 27 for monotherapy and 60 for polytherapy, with the most frequent seizures (more than 1 seizure per month) observed for individuals on polytherapy.

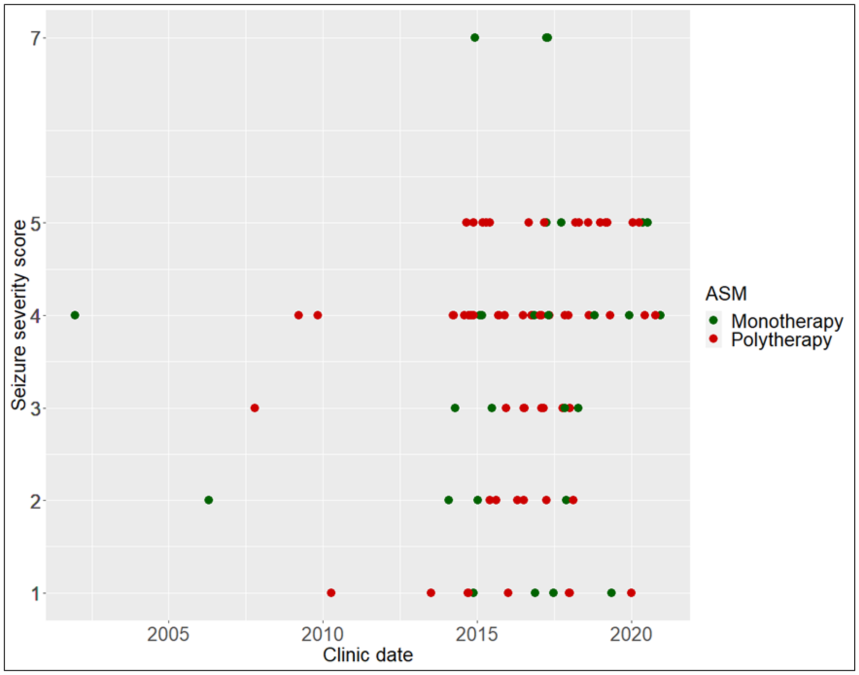


Figure 6.15: *ASM monotherapy and polytherapy and seizure severity score for GTCS.*

Investigating the monotherapy prescriptions in more detail, out of the 122 dated records for linked ASM and SSS, the largest group was for levetiracetam, with 43 records, followed by lamotrigine with 30, valproate with 26, carbamazepine with 12, topiramate with 4, eslicarbazepine 3, and one each for gabapentin, phenytoin, and brivaracetam.

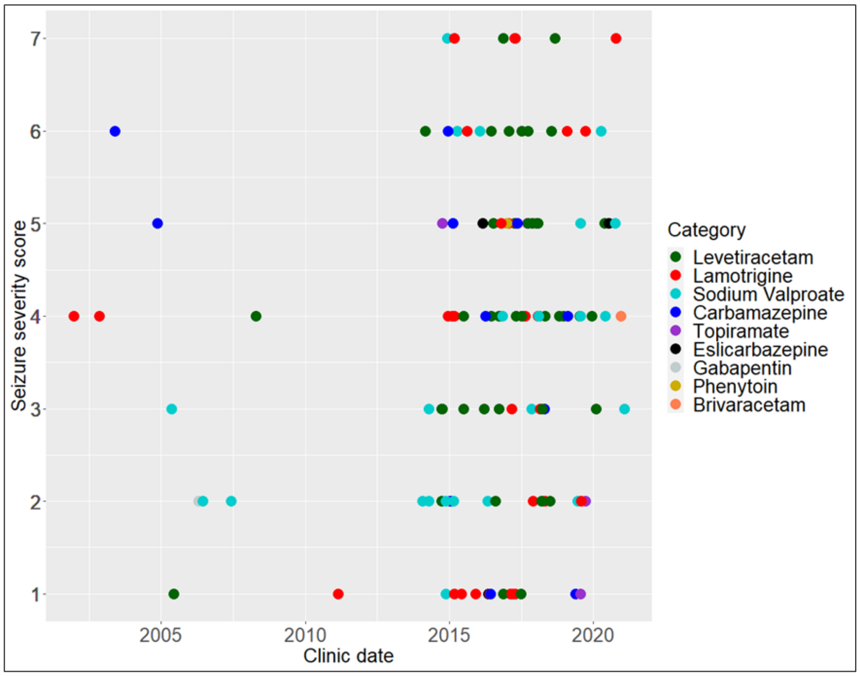


Figure 6.16: *ASM monotherapy and Seizure Severity Score by specific ASM.*

It is difficult to notice a specific pattern, but it seems that there are more cases of the ‘less than 1 per year ’seizure frequency for records linked to sodium valproate (9 out of all sodium valproate prescriptions) and levetiracetam (7 out of 43). At the same time both of these and lamotrigine are shown in the records for the most frequent seizures.

A similar analysis was carried out for the polytherapy records by the most frequently grouped ASMs. There were 119 records of SSS with two ASM prescriptions. These contained 32 records of levetiracetam with lamotrigine, 15 of carbamazepine with levetiracetam, 11 of topiramate with another ASM not already grouped, nine each of sodium valproate with lamotrigine, eslicarbazepine with another ASM not already grouped, and carbamazepine with clobazam, seven each of sodium valproate with levetiracetam, and brivaracetam with carbamazepine, and nine other combinations (not shown on the plot), Fig. 6.17.

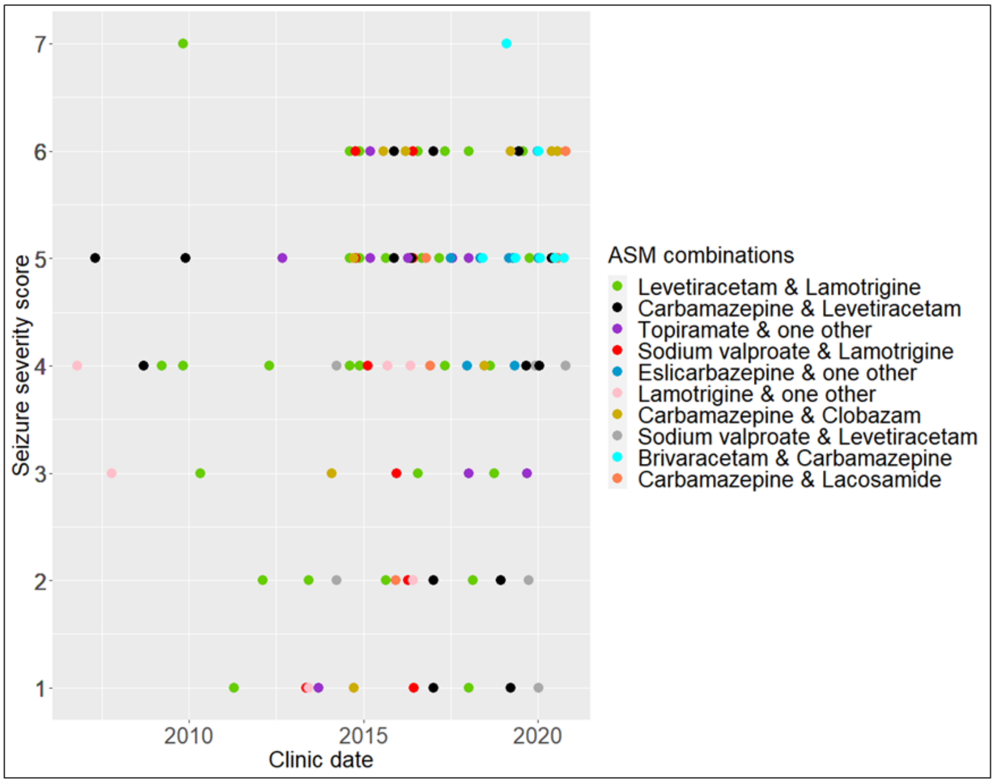


Figure 6.17: *ASM polytherapy and Seizure Severity Score by combinations of two ASMs. topiramate was used with lamotrigine, lacosamide, carbamazepine, levetiracetam, or eslicarbazepine*

There were also 45 records of three ASMs and 16 of four. The most common combinations are shown in fig 6.18, with 15 of levetiracetam with lamotrigine, and clobazam or clonazepam; 13 of levetiracetam with lamotrigine, combined with other two ASMs; nine of sodium valproate with levetiracetam, combined with carbamazepine, topiramate, or lamotrigine; five of lacosamide with topiramate, sodium valproate, or carbamazepine, and one other ASM; four of eslicarbazepine with topiramate, levetiracetam, or zonisamide and one other ASM; four of carbamazepine in combination with two other ASMs not present in the previous combinations; and another five of three ASMs, and three of four ASMs combinations.

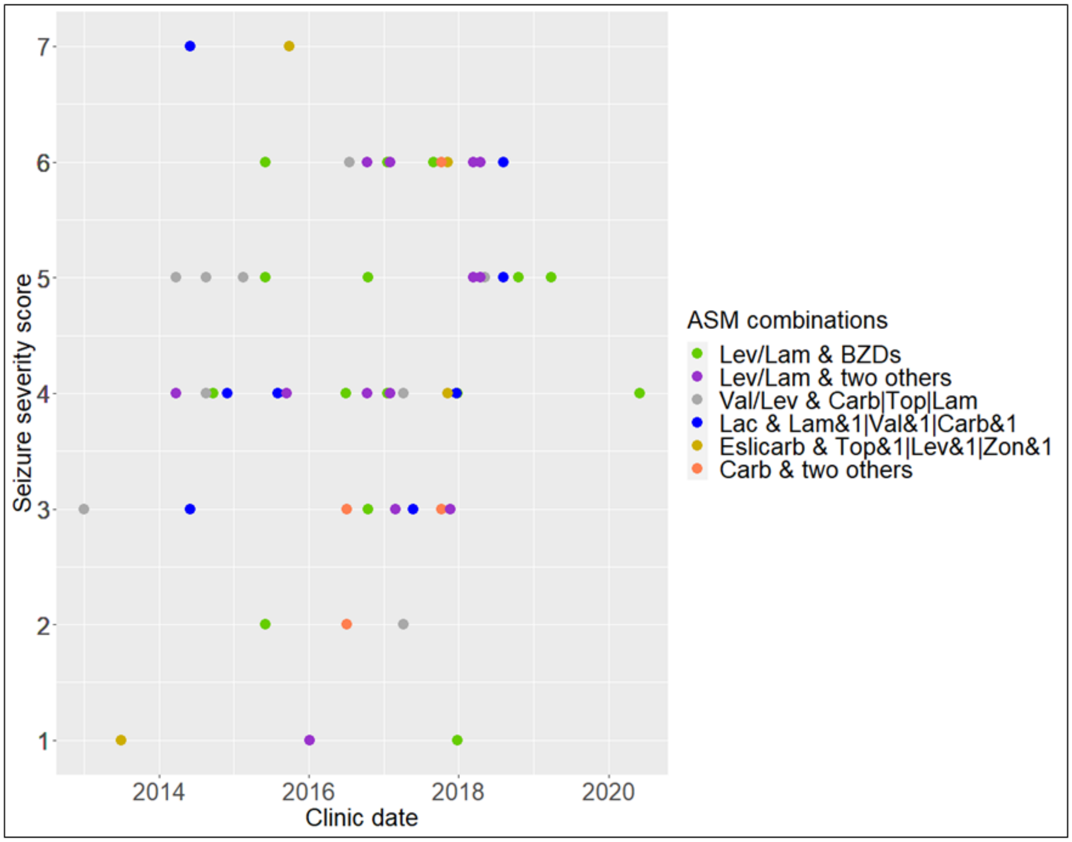


Figure 6.18: *ASM polytherapy and seizure severity score by combinations of three or four ASMs. Lev/Lam & BZD = LEV and LMT with clobazam or clonazepam; Lev/Lam & two others = LEV and LMT in combination with other two ASMs; Val/Lev & Carb|Top|Lam = VPA and LEV in combination with CBZ, TPM, or LMT; Lac & Lam & 1| Val&1|Carb&1 = LCM in combination with LMT and 1 other, or VPA and 1 other, or CBZ and one other ASM; Eslicarb & Top & 1|Lev&1|Zon & 1 = ESL in a combination of TPM and one other, or with LEV and one other, or with ZNS and one other ASM; Carb & two others = CBZ and other two ASMs not already included in the previous combinations.*

It is very difficult to see any clear pattern between the SSS and the polytherapy groupings. It is striking that even on three ASMs some individuals experience a very high frequency of seizures. Some of the combinations of three may be due to the changeover of the medication but all should reflect the current combination, not a commencement of a new drug. The records presented disregard the type of epilepsy, seizure type, and gender. The splitting of the cohort would produced very small groupings, but it might be that a more clear pattern would emerge.

6.4 Chapter summary

This chapter presented the output produced by the ExECT pipeline from a range of annotations on clinic letters from the Epi25 study. It described the validation of these outputs against the SNB database extract, presented how they could be linked in a longitudinal analysis by individual, and investigated the linkage of seizure frequency and ASM for the entire cohort.

- The output tables were linked to the Record Date produced by IDEx and processed to create SYSTEM_ID for each dated record, allowing for the linkage of different annotation types by individual and record date.
- The diagnosis annotations were processed to extract diagnosis by epilepsy type from the epilepsy diagnosis and seizure types stated in the letters. These were then validated against the SNB records achieving very high scores. The analysis of a small number of errors revealed that for some cases the letters did not contain the information required, i.e. no annotations could be made by ExECT, and that in two cases the diagnosis has changed since the biobank database was updated.
- The output for onset annotations could not be validated against the biobank, as onset is not routinely repeated in clinic letters. For just under a quarter of the individuals in the cohort the onset information was extracted, and these records were compared to the SNB dataset. Most of the records were fully matched, some were partial matches, and those that were not matched on inspection were shown to be correctly extracted by the pipeline, but the expression of age within the letters was not exactly the same as that recorded in the biobank.
- For the patient history annotations, only febrile seizures, depression, and dissociative seizures were compared with the SNB records. From these, only the negation status for the small number of febrile seizures matched by individual was validated and all records were matched. For the two comorbidities selected, a very small number extracted by ExECT was also recorded in the SNB database.
- The validation of investigation results (EEG and MRI) was performed on grouped (normal, abnormal) results per individual. Only the EEG abnormal scores were

comparable with those achieved by the validation on the 100-letter set. This led to a detailed review of the annotations and the letters for individuals with the false negative results. Most of the EEG annotations produced by the pipeline were correct, with the exception of the false positives for abnormal results, where only a half were identified as true false positives. The review of the MRI annotations identified a larger proportion of true errors, but most of them were less than half of the number shown in the validation. A range of reasons can be put forward as an explanation of this discrepancy, the availability of results during the data collection for the biobank, access to investigation reports in paper notes, limited number of letters within the document cohort without the full history, but also more recent documents with the latest result. Though, some gaps in the results gazetteers were also identified.

- The detailed precessing of seizure frequency annotations, resulting in the construction of seizure severity scores and a similar process for the prescription annotations, producing a total daily dose for each ASM, allowed for the linkage of dated records for the two variables for each individual.
- Using the linkage of the SSS and ASM, records of three individuals over the study period were reviewed as an illustration of how the output from the pipeline could be used as a clinical resource.
- The daily ASM dose and the SSS were linked for monotherapy and polytherapy and by the most commonly grouped ASMs. No clear patters emerged from this analysis.

Chapter 7

NLP WITHIN SAIL DATABANK AND GENETIC DATA LINKAGE

This chapter describes how clinical data extracted with an NLP pipeline can be used to enrich information within the SAIL databank. A path-finder study linking genetic data and routinely-collected data within the SAIL databank is used to illustrate this process. Seizure frequency information produced by the ExECT pipeline is added to the analysis to provide an additional level of detail not otherwise available.

7.1 Epilepsy genetics data linkage study

The Epilepsy genetics data linkage study was a pathfinder project aiming at establishing a pipeline for linking genetic sequence data with routinely collected health data within the SAIL databank. The study used exome data for a cohort of individuals with epilepsy who donated samples to the SNB. Sequencing was performed as part of the Epi25 Collaborative between 2016 – 2018. The resulting Variant Call Format (VCF) files were uploaded and annotated within the SAIL gateway, and linked to two epilepsy outcome measures, unscheduled admissions with a diagnosis of epilepsy and ASM polytherapy. This process is fully described in the Method

Chapter with fig. 2.10 illustrating the uploading and annotating of VCF files and fig.2.14 giving the details of data linkage and analysis.

7.1.1 SNB cohort within the SAIL databank

From the original 111 VCF files, 107 were linked to GP registered patients within the SAIL databank, and this group was used in the analysis of unscheduled hospital admissions. 105 individuals had a primary care ASM prescription record and this group was used in the analysis of polytherapy. 104 people had epilepsy classification information extracted from the uploaded SNB clinical dataset, and a separate analysis for the two outcome measures, by epilepsy type, was carried out for these individuals, Table 7.1. The earliest GP event date relating to epilepsy (diagnosis or ASM prescription) was 2000 and the latest 2019, giving a study window of 19 years (mean=12 years, range 2–19 years).

	All	**Admissions**		**Polytherapy***		**Seizure frequency****	
		No admissions	Admissions	Monotherapy	Polytherapy	>1 year seizure freedom	<1 year seizure freedom
Total	107	79 (74%)	28 (26%)	32 (31%)	73 (69%)	10 (10%)	87 (90%)
Mean age/years	41	41	40	37	42	41	41
Male	47 (44%)	38 (48%)	9 (32%)	14 (44%)	32 (44%)	4 (40%)	36 (41%)
Female	60 (56%)	41 (52%)	19 (68%)	18 (56%)	41 (56%)	6 (60%)	51 (59%)
Focal Epilepsy§	73 (70%)	54 (71%)	19 (68%)	18 (58%)	54 (76%)	5 (56%)	61 (71%)
Generalised Epilepsy§	31 (30%)	22 (29%)	9 (32%)	13 (42%)	17 (24%)	4 (44%)	24 (28%)

Table 7.1: *Total study and outcome group characteristics. Age was calculated for the end of the study.* * *There was no record of anti-seizure medication (ASM) prescriptions for two individuals and so there were a total of 105 individuals in the polytherapy outcomes group.* ** *Seizure frequency scores were extracted for 100 individuals (see method) but only 97 could be liked within SAIL to sex/age data, 94 had diagnostic information in the Swansea Neurology Biobank (SNB) (66 focal and 28 generalised)* § *Epilepsy classification was not available for 3 individuals.*

7.1.2 Rare variants burden analysis

Rare and potentially damaging variants were defined as those with Allele frequency (AF)<0.001 as per gnomAD exome collection (v2.1.1), [197] Combined Annotation-Dependent Depletion (CADD) score, (section 2.6.8 on page 65) of ≥15, [176] and not present more than twice within the study cohort. The initial stage of filtering for AF<0.001, CADD_PHRED ≥ 15 produced 4436 variants, from 28 to 81 per person, which on removing those present more than twice was reduced to 850 qualifying variants, 1–48 per person (780 variants, 710 occurring once and 70 twice). The SQL script for selecting the qualifying variants is given in Appendix **??** and the resulting numbers of variants are shown in fig7.1

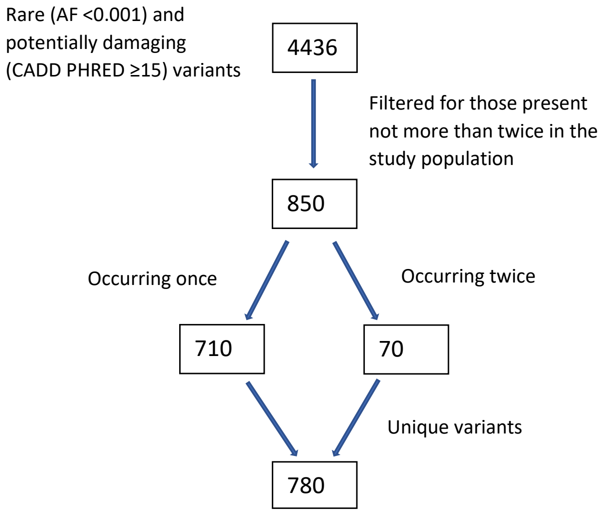


Figure 7.1: *Summary of the filtering process for rare and potentially damaging variants that are not present more than twice in the study population, giving the numbers of variants at each stage.*

7.1.3 Exome-wide burden

Two methods of calculating an individual's exome-wide burden were used. Cumulative scores, which was a sum of CADD PHRED scores for all rare and damaging variants [167], and the overall number of rare variants for each individual. [217] Cumulative scores ranged from 23 to 1,245.59, with the number of rare variants ranging from 1 to 48 per person. A table of cumulative scores was created to use in the analysis and added to the project, and the script used in shown in Fig.7.2. The filtering process for rare and potentially damaging variants for the overall number of variants per individual and for linkage to the epilepsy and metabolism/transporter genes is given in the Appendix D.3 on page 242

```
 -- needed to connect to SAIL project
install.packages("SAILDBUtils")
library(dplyr) -- package to be used

%library(RODBC)
%source("/shared/0000 - Analysts Shared Resources/R/sail_login.r")    --connecting to SAIL

-- filtering the VCF table for a selection of needed fields, converting CADD_PHRED into decimals
vcf <- sqlQuery(channel,
"select * from
(select * from
(SELECT DISTINCT VCF_FILE_PE, AF, CLNDN, FUNC_REFGENE, CADD_PHRED,CAST(CADD_PHRED AS DECIMAL (10,3)) AS CADD_PHRED_DEC, case when AF LIKE '%e%' then '0.012345' else AF end as AF_temp FROM SAIL0661V.EPI25VCF
WHERE AF not in ('-','.')
fetch first 9000000 rows only   -- more than needed but if omitted seems to pick up a non-flot entry
)
where float(AF_TEMP) < 0.001 --filtering for AF less than 0.001
AND CADD_PHRED_DEC >=15)" -- filtering for CADD score at 15 or more
)
-- grouping by individual (VCF_FILE_PE) and summing up the scores for each person
burdens <- vcf %>%
group_by(VCF_FILE_PE) %>%
summarise(burden = sum(CADD_PHRED))
-- removing an older table from the project schema
sqlQuery(channel,"drop table SAILW0661V.BURDENS4")
-- adding the new table to the project schema
sqlSave(channel,burdens,"SAILW0661V.BURDENS4", rownames=FALSE)
```

Figure 7.2: *SQL script creating Cumulative Burden table – filtering rare and potentially damaging variants, summing up the scores for each person, and creating a table which is uploaded to the project schema.*

7.1.4 Gene-based variant burden

For gene-based analysis rare and damaging variants selected through filtering were further filtered for variants in genes: (i) associated with epilepsy i.e., epilepsy genes, neurodevelopment-associated epilepsy genes and epilepsy-related genes [2] [1, 167] and (ii) associated with drug metabolism and/or drug transporters, i.e. genes encoding phase-1 drug metabolising enzymes and genes encoding drug transporters. [3]. Full list of epilepsy and drug metabolism / transporter genes is given in Appendix D.2

Residual Variation Intolerance Scores Residual Variation Intolerance Scores (RVIS) are a measure of gene's tolerance of functional variations. RVIS ranks genes according to 'whether they have more or less common functional genetic variation relative to the genome wide expectation given the amount of apparently neutral variation the gene has'. [218] Genes are ranked from the most tolerant (positive scores) to the least tolerant (negative scores). A table containing RVIS expressed as percentiles [218] was uploaded to the SAIL gateway and used for ranking genes identified following filtration in each of the cohort subgroups.

7.1.5 Genetic burden and Unscheduled Hospital Admissions

Primary care Read codes were used to identify epilepsy diagnoses (Read v2 "F25"), anti-seizure medication (ASM) prescriptions (Read v2 "dn%"), and dates within the WLGP [118]. To identify unscheduled hospital admissions for epilepsy, PEDW admission method codes "21" and "29" and the ICD10 G40 diagnosis were used. To select admissions whilst on anti-seizure therapy further filtering was carried out for admissions following ASM commencement date. Two groups were identified by this selection, 79 individuals without a history of unscheduled hospital admission for epilepsy and 28 with a history of such admissions.
Genetic burden as defined in 7.1.4 was then assessed for each of the groups, fig. 7.3. The Wilcoxon rank sum test was used to compare the groups and no difference was

identified, with p-value = 0.82 for the cumulative score comparison and p-value = 0.85 for the number of variants.

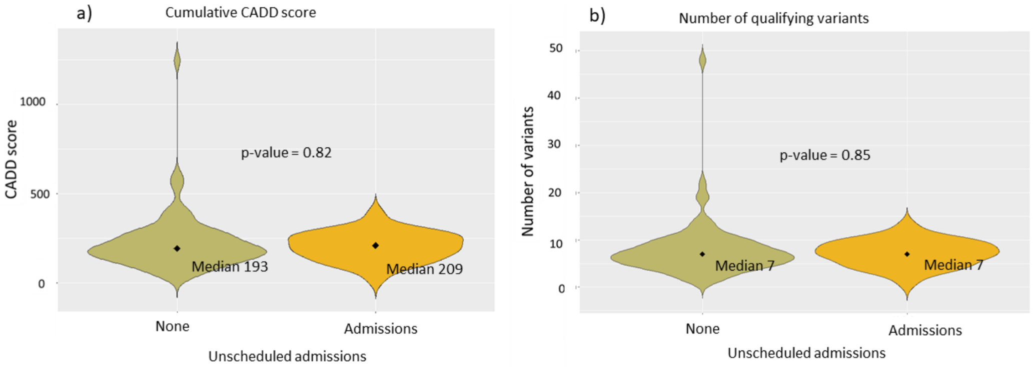


Figure 7.3: *Violin plots of people with epilepsy and unscheduled hospital admissions with a diagnosis of epilepsy while on ASM therapy, individuals without a history of unscheduled hospital admissions (green) and people with epilepsy with a history of such admissions (yellow), in terms of: (a) cumulative CADD score for qualifying variants (b) number of qualifying variants. The width of the plots represents the probability density and medians are shown on the graphs.*

562 (74.1%) and 168 (22.2%) unique qualifying variants in 501 (72.2%) and 145 (20.9%) genes were identified in the no admissions and the admissions groups respectively. After filtering for epilepsy and drug associated genes, variants exclusively present in each of the groups were identified. Figure 7.4 shows the numbers of variants and genes in each of the groups, and separately, the epilepsy and drug metabolism/transporter genes affected. The number of individuals in each of the groups was too small for statistical testing [219] and not allowed for reporting under the SAIL guidance on disclosure. [220]

Variants in two genes associated with epilepsy, in *CACNA1C* and *KCNQ1* were exclusively present in the admissions group (both affecting fewer than 5 individuals). Qualifying variants were seen in three drug metabolism and transporter genes, with *CYP2D6* being present exclusively in the no admissions group. *CACNA1C* and *KCNQ1*are amongst the top 3% of genes intolerant to damaging variants (RVIS score) with *CYP2D6* being in the most tolerant group, 96%.

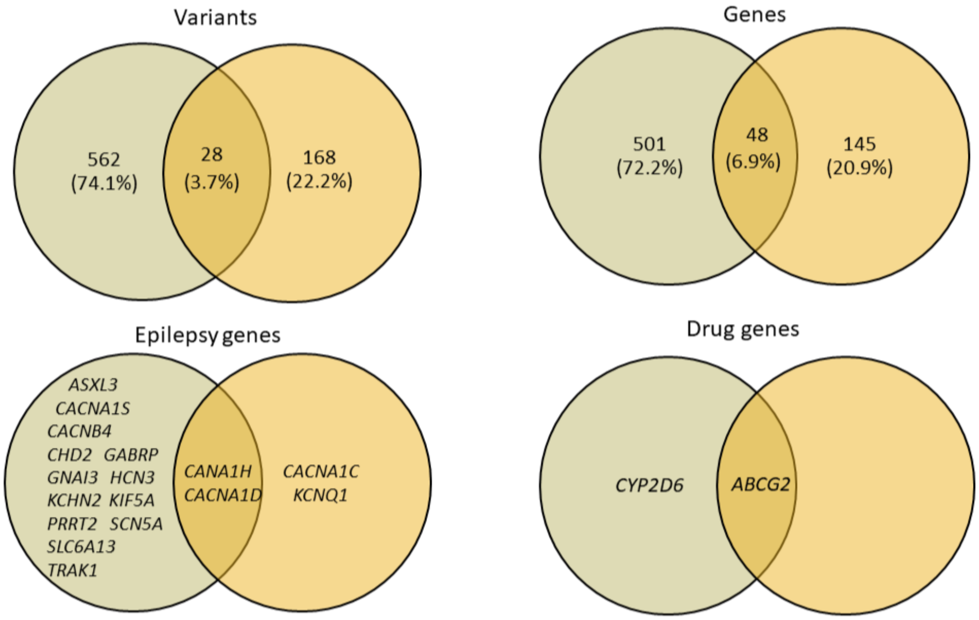


Figure 7.4: *Venn diagrams of gene-based burden for qualifying variants in the no admissions (green) and the admissions (yellow) groups. Top row shows the numbers of unique and shared variants in each of the groups (left) and the number of unique and shared genes (right). Bottom row gives genes with unique and shared variants in the no admissions and the admissions group,for epilepsy genes (left) and for drug metabolism and transporter genes (right)*

7.1.6 Genetic burden and GP records of ASM

ASM prescription information from GP event data was available for 105 out of the 107 registered individuals as 2 had no ASM prescription records, Table 7.1. Two distinct groups were identified, 32 people who were prescribed only one ASM (monotherapy) versus 73 who were prescribed two or more ASMs for at least six months (polytherapy). It was assumed that 6 months was a long enough period to disregard medication change due to adverse drug reaction, treating the change as an indication of poor seizure control, although for some drugs this period may be longer [221].

Genetic burden as described in **??** and 7.1.4 was analysed for the two groups. The Wilcoxon rank sum test was used to compare the groups and no difference was identified, with p-value = 0.41 for the cumulative score comparison and p-value = 0.37 for the number of variants.

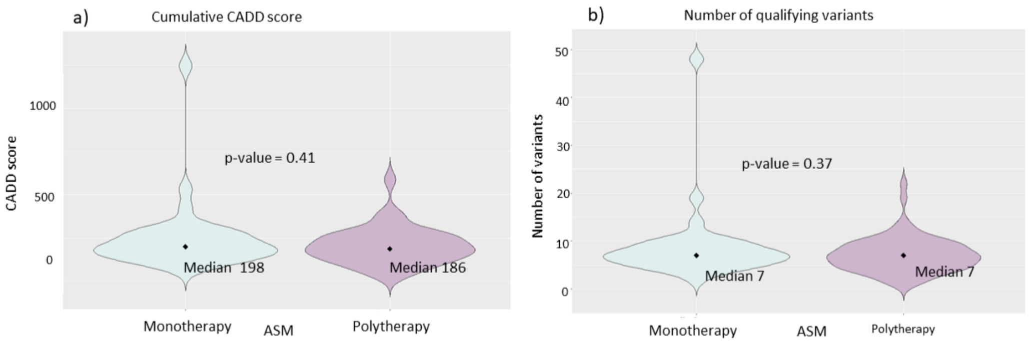


Figure 7.5: *Violin plots for individuals with epilepsy on anti-seizure medication (ASM), monotherapy (light blue) compared to those on polytherapy (purple), in terms of: a) cumulative CADD score for qualifying variants; b) number of qualifying variants. The width of the plots represents the probability density and medians are shown on the graphs.*

249 (33.6%) and 469 (63.3%) unique qualifying variants in 214 (31.6%) and 415 (61.2%) genes were identified in the monotherapy and polytherapy groups, respectively. Following filtering for epilepsy and drug associated genes, sets of variants exclusively present in each of the groups were identified. They affected a small number of individuals, and, as in the case of the admissions group, section 7.1.5, no statistical test was applied or data presented, hence fig. 7.6 shows only cumulative CADD scores and number of qualifying variants.

There were a number of variants for epilepsy associated genes exclusive to both groups, with the polytherapy group having more genes from amongst the top 3% of genes intolerant to damaging variants (RVIS). There were no specific drug metabolism and transporter genes with qualifying variants that were uniquely associated with polytherapy. Two intolerant genes *CACNA1C* and *KCNQ1* (in the top 3% of intolerant genes) were present in both unscheduled admissions and polytherapy groups.

7.1.7 Genetic burden analysis by epilepsy type

Detailed epilepsy type information (epilepsy and seizure type) was extracted from the SNB database and uploaded to the SAIL Databank following the standard split file approach. [220] This process is described in more detail in section 2.1.3 on page

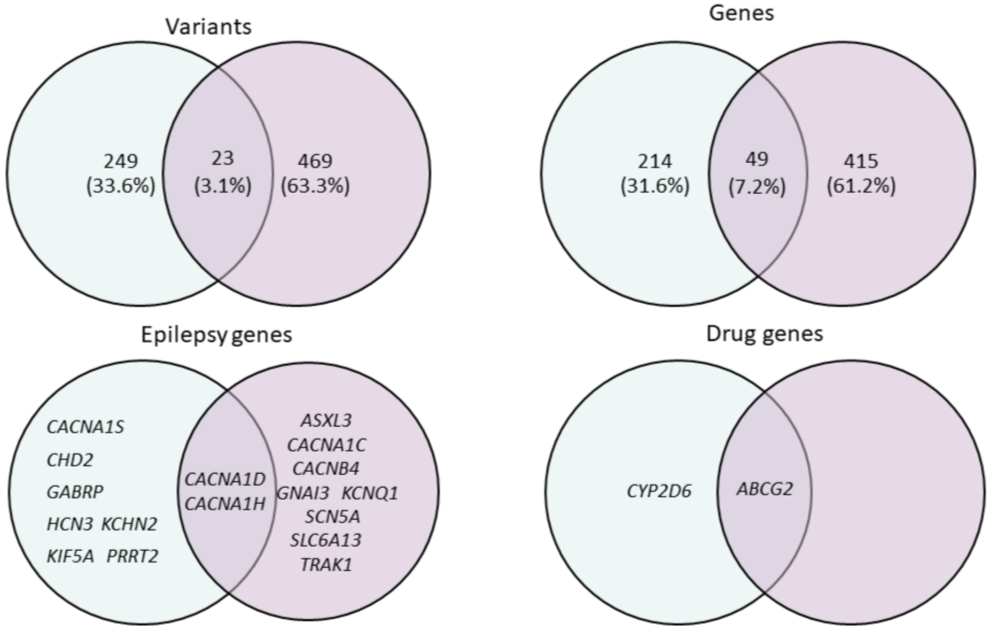


Figure 7.6: *Venn diagrams of qualifying variants (top left), genes (top right), epilepsy genes(bottom left) and drug metabolism and transporter genes (bottom right) in individuals on ASM monotherapy versus those on polytherapy.*

40. The SNB clinical dataset was linked to GP registrations and a single table with the biobank clinical data for registered and sequenced individuals was created. During this process it was noted that for two individuals biobank diagnosis was recorded as ‘Features of focal and generalised’ which, following a review of seizure type extracted, was identified as 1 case of focal and 1 case of generalised epilepsy and recoded after linkage and filtering. This table was then linked to Cumulative scores (described in 7.1.3 on page 162) and to PEDW admissions and GP event data to investigate exome-wide cumulative score by epilepsy type for unscheduled admissions history and ASM therapy. Similar process was carried out to compare the numbers of qualifying variants for each of the groups. Epilepsy diagnosis, 73 (70%) focal and 31 (30%) generalised, was confirmed for 104 linked individuals, fig. 7.1

Genetic burden as cumulative score and number of qualifying variants were compared for different epilepsy types. Generalised epilepsy showed a higher burden in both measures, with p-value = 0.02 for cumulative score and 0.02 for the numbers of variants. The test could not produce the exact p-values due to ties and the sample size of < 50, so a normal approximation value was returned fig 7.7

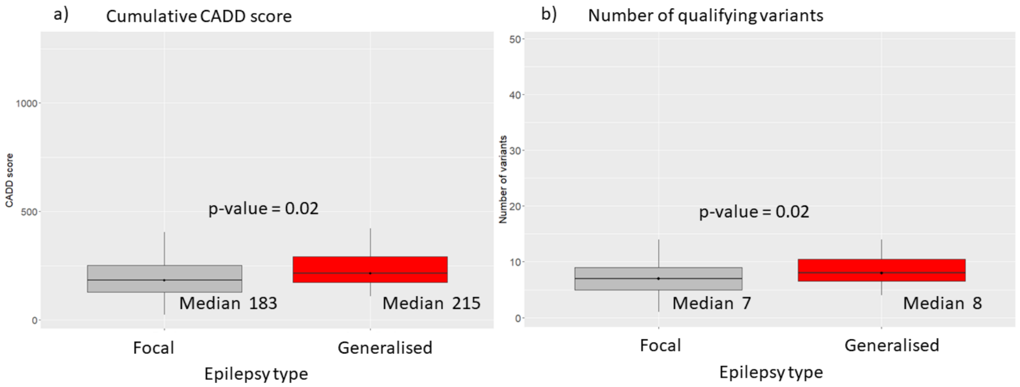


Figure 7.7: *Box plots of people with focal (non-acquired) epilepsy (grey) and generalised epilepsy (red) compared in terms of: (a) cumulative CADD score for qualifying variants (b) number of qualifying variants.*

21 variants in 16 epilepsy genes were found in the linked biobank cohort, 9 in focal and 11 in generalised epilepsy, with some exclusively present in each of the groups, as shown in fig. 7.8. 4 variants of 2 drug metabolism and transportation genes were present, 3 in *ABCG2*, 1 focal and 2 generalised epilepsy, and 1 in *CYP2D6* in generalised epilepsy patient.

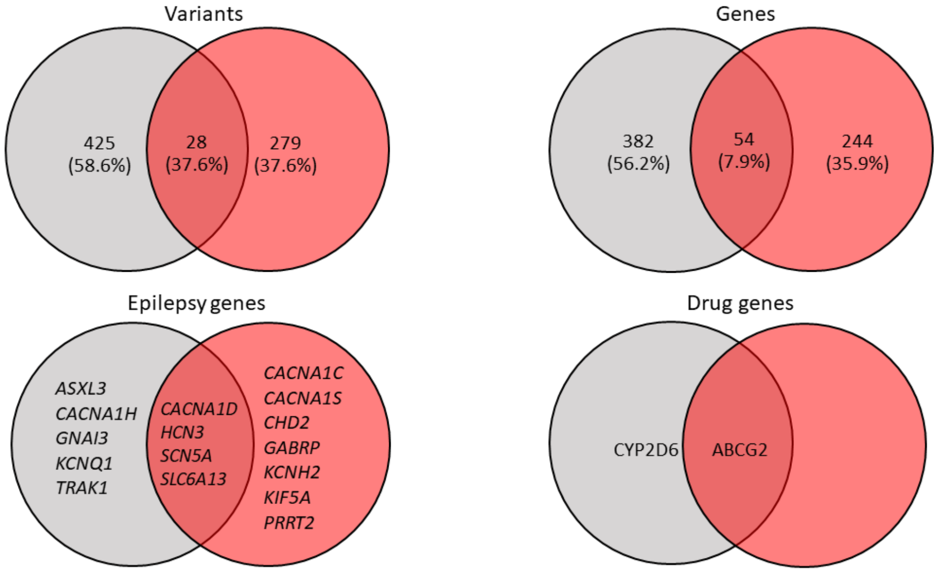


Figure 7.8: *Venn diagrams of qualifying variants (top left), genes (top right), epilepsy genes(bottom left) and drug metabolism and transporter genes (bottom right) in individuals with focal epilepsy (grey) and generalised epilepsy (red)*

Epilepsy type and unscheduled hospital admission

Analysis by epilepsy type showed a difference in exome-wide cumulative CADD score in the generalised epilepsy group, suggesting a greater burden in people who did not have an unscheduled admission, median = 262, than in people who were admitted, median = 176, with p-value = $0.04 < 0.05$. This difference was not evident in the analysis of the number of variants, where the median values were respectively 9.46 and 2.40, with p-value = 0.06. It has to be noted that this analysis was based on very small groups, 9 individuals with and 22 without admissions. Also, some individuals within the no admissions group had very high cumulative scores associated with specific syndromes, which seemed to influence the result.

No separate analysis was carried out for gene-based analysis by epilepsy type for the two measures used as the numbers of individuals in each group and the number of genes was very small.

Epilepsy type and ASM therapy

7.2 Linking NLP output to the genetic study within SAIL gateway

Linking the NLP output to routinely collected data is one of the main objectives of developing this NLP applicaton. It aims to enrich health information available from the EHR datasets to provide intricate clinical details not otherwise accessible.

7.2.1 Per person Epi25 cohort NLP pipeline outputs

Outputs from ExECT supplemented with the record date produced by IDEx were uploaded to the SAIL gateway in csv format following standard procedure for linking data within SAIL Original biobank numbers were encrypted creating SYSTEM_ID_PE for linkage and Db2 data tables were created and added to the project by SAIL technical team. For convenience, before data upload, smaller ExECT output files were combined, so that from the original output of nine, four data tables were created.

ExECT output	File for SAIL upload
Diagnosis, Investigations	EPI25DIAGINVEST_0661_20211111.csv
Patient History, Onset, Cause, When Diagnosed, Birth History	EPI25PHONSETCWDBH_0661_20211111.csv
Prescriptions	EPI25PRESC_0661_20211111.csv
Seizure Frequency	EPI25SF_0661_20211111.csv

Table 7.2: *ExECT output groupings for upload into the SAIL genetic data linkage project.*

Apart from specific subject annotation fields, each file contained SYSTEM_ID_PE, DATEREC (created by IDEx), and DOC, an additional field created from the original document reference, which was substituted with a random number, and an existing sequence number within the set, so that the document reference could be maintained for analysis using an annotation start point, such as prescription. '*Annotation*'field was also added to identify the original output file in the amalgamated tables.

7.2.2 Seizure frequency ExECT output linkage

Seizure frequency was one of the main ExECT outputs of interest for the epilepsy genetics linkage project, as it could provide very detailed and up-to-date information on an individual's seizure control status. In order to extract seizure frequency information which could be compared between different individuals, ExECT output had to be processed, as described in section 6.3.1. As previously (see figure 6.5) Seizure Severity Scores were used to compare seizure status. If more that one score was present in a singe dated record, the highest score was used, as not to underestimate the severity reported.
Seizure frequency processing was done in R Studio (R-version 4.1.1) with the dplyr package and SAILButtils for Db2 connection and table upload to the project schema.

Processed seizure frequency output as per-person dated SSS for different seizure types (CUIs) was added to the project schema as SAILw0661v.SFSEVERITYCUI. 100 individuals had a seizure frequency record resulting in a score. Records that were excluded from the analysis were those where seizure frequency was reported as 'increased' or 'decreased' or those without a specified time period, date, or point in time. SYSTEM_ID_PE was used for linkage to genetic data tables, cumulative CADD score, number of rare and damaging variants, and to Epi25 clinical data. In order to compare individuals as closely in time as possible the most recent seizure frequency (severity score) within the study period was selected. If more that one score was present in a single dated record, a highest score was used, as not to underestimate the frequency reported. Seizure type was disregarded in this analysis but it is something that is going to be included in later work. The process of linkage is shown in fig 7.9.

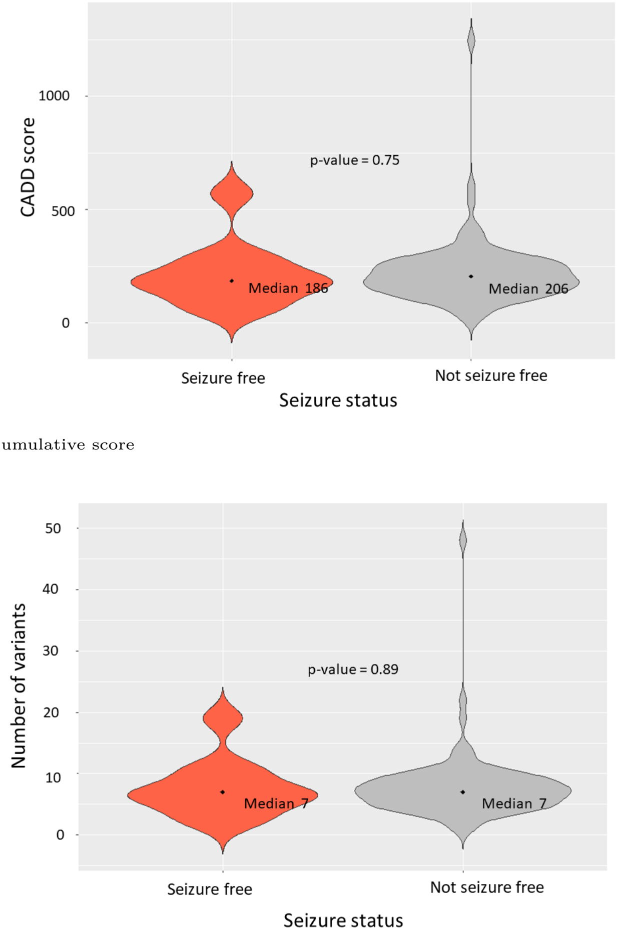
```sql
CREATE temp TABLE SAILw0661V.Epi25_SF AS -creating table of per person latest seizure scores
SELECT SYSTEM_ID_PE , MAX(FREQSEVERITY)AS MAXSF FROM  -- extracting the highest severity score
(SELECT SYSTEM_ID_PE, FREQSEVERITY FROM
(SELECT b.SYSTEM_ID_PE , a.MaxDate , b.FREQSEVERITY FROM
(SELECT SYSTEM_ID_PE, MAX(DATEREC) AS MaxDate -- extracting the  most recent date
FROM SAILw0661v.SFSEVERITYCUI  -- seizure frequency table
GROUP BY SYSTEM_ID_PE)a
INNER JOIN SAILw0661v.SFSEVERITYCUI b
ON a.SYSTEM_ID_PE = b.SYSTEM_ID_PE))
GROUP BY SYSTEM_ID_PE

SELECT * FROM SAILW0661V.Epi25_SF a --seizure severity score table created above
JOIN SAILw0661V.BURDENSRARE  b   --table with rare variants and CADD score >=15 and not present more than twice
ON a.SYSTEM_ID_PE = b.VCF_FILE_PE
-- linking seizure severity scores to the number of rare variants with CADD >=15
SELECT * FROM SAILW0661V.Epi25_SF a
join
(SELECT "VCF_FILE_PE", COUNT(*) AS NO_OF_VARIANTS
FROM SAILW0661V.EPI25_RARE_VARIANS_CADD  --rare variant table
GROUP BY VCF_FILE_PE
ORDER BY VCF_FILE_PE) b
ON a.SYSTEM_ID_PE = b.VCF_FILE_PE

-- linking to epilepsy genes using SAILw0661V.VARIANTS_EPI_GENE -- uploaded epilepsy genes table
SELECT * FROM SAILW0661V.Epi25_SF a
join
SAILw0661V.VARIANTS_EPI_GENE b --
ON a.SYSTEM_ID_PE = b.VCF_FILE_PE

-- linking to drug genes using SAILw0661V.VARIANTS_DRUG_GENE -- table based on uploaded drug genes list
SELECT * FROM SAILW0661V.Epi25_SF a
join
SAILw0661V.VARIANTS_DRUG_GENE b -- only 4 drug variants selected in rare variants in our group
ON a.SYSTEM_ID_PE = b.VCF_FILE_PE


```

Figure 7.9: *Seizure frequency (SSS) linkage within the SAIL genetic data linkage project.*

Two separate groupings were constructed to perform genetic burden analysis. One for individuals who were seizure free for more than a year (n = 10) versus those with active seizures (n = 90), another for those seizure free for more than a year versus those with at least weekly seizures (n = 46).

Seizure frequency and exome-wide burden

Exome-wide burden measured as (i) cumulative CADD score and (ii) number of qualifying variants, as defined in 7.1.2 was analysed for the two groups in turn. There was no significant difference between the groups. 62 (6.2 per person) and 684 (7.6 per person) unique qualifying variants in 62 and 586 genes were identified in the seizure free and not seizure free groups, respectively. There was no difference in the genetic burden between the groups, with the p-value = 0.75 and 0.89 for the two tests in the seizure free versus non seizure free analysis,Fig. 7.10. Further testing, comparing seizure free individuals to those experiencing at least weekly seizures did not produce different results.



(a) Cumulative score

(b) Number of variants

Figure 7.10: *Seizure frequency and genetic burden: Seizure free for at least 1 year vs Not seizure free individuals, more than 1 seizure per year.*

Gene-based burden and seizure frequency

Out of the 22 individuals identified previously as having qualifying variants in genes associated with epilepsy, 19 were linked to the seizure frequency output. Of those, only one was seizure free for more than 1 year, which made any caparison between the groups impossible. There were 3 individuals identified as having drug metabolism and transporter gene variants, of which one was seizure free (*CYP2D6*) and 2 had active epilepsy *(ABCG2)*.

One epilepsy associated gene was found in the seizure free group and 15 in the not seizure free group, including *CACNA1C, KCNQ1*, and two less tolerant genes (top 4%), *SCN5A, TRAK1*, also present in the polytherapy group. Variants of *CHD2* and *KCNH2* were present exclusively in the not seizure free group, both from the top 4% of the least tolerant genes. A small number of variants of *ABCG2* were present exclusively in the not seizure free group (not shown).

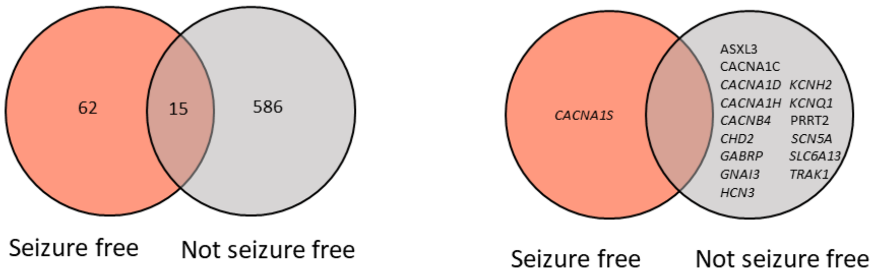


Figure 7.11: *Seizure frequency, genes associated with epilepsy, and the number of qualifying variants*
Epilepsy

Genetic burden and Seizure frequency by epilepsy type

The seizure frequency output was linked to genetic data for different epilepsy types as extracted from clinical data from the SNB, primarily to test the method as the number of seizure free individuals was very small. No difference was observed in the genetic burden between people who were seizure free for more than 1 year and those not seizure free for different epilepsy types.

7.3 Chapter summary

- The aim of the genetic data linkage study was to establish a pathway for uploading, annotating, and liking genetic variant data to the routinely collected datasets within the SAIL gateway.
- The linkage process was described in detail and the variants selected were discussed. Genetic data for 107 individuals with epilepsy was linked to electronic health records, 26% had unscheduled hospital admissions and 70% were prescribed anti-seizure medication polytherapy.
- There was no significant difference between the outcome groups in terms of the exome-wide and gene-based burden of rare and deleterious genetic variants.
- The seizure frequency output from ExECT v2 was linked to the study. No difference was observed between the seizure free and not seizure free groups in terms of the exome-wide and gene-based burden of rare and deleterious variants.
- By linking genetic data and the outputs from the NLP processing of clinic letters within the SAIL databank, this study established a novel method to health records linkage.

Chapter 8

DISCUSSION

This chapter discusses the results presented in chapters 3-6 ,

8.1 Gold standard annotation

The creation of a gold standard annotation set is essential for the development and validation of an information extraction application. [185] This is a long and complex process, requiring the creation of annotation guidelines, training of annotators, and reviews of annotations to arrive at the most accurately annotated set of documents. [156, 222] Apart from being used for the algorithm's validation, the gold standard may provide a benchmark for the results achieved if the inter annotator agreement (IAA) is calculated. The aim of this part of the project was the creation of a gold standard annotation set for epilepsy clinic letters. The results presented in chapter 3 are discussed here.

8.1.1 Annotation guidelines

Annotation guidelines are developed to train and aid the annotators in the annotation tasks to guarantee consistency of standard and to minimise errors. The development of the guidelines for this project followed a well documented iterative approach. [154, 156, 223] The set of entities to be annotated and the features were agreed by the team at the start of the project and contained all the previously extracted variables from the original ExECT pipeline project and a set of new concepts. The first draft of the guidelines contained the definitions of these entities and features together with a small number of examples. The approach was that the guidelines would evolve during the trial annotations in response to the annotators' needs. The annotators were also asked to contribute to the list of concepts if they felt that terms should be added.

The development was therefore partly led by the annotators, who requested clarifications, lists, and examples, and partly by the algorithm developers, who had to ensure that the format of the annotations created was in line with that produced by the algorithm, to allow for validation.

For the named entities, the approach was similar to that used in the CLEF project corpus, it was as important to clearly state what should as what should not be annotated. [224] For example, much time was spent by the team deciding on the annotation of non-specific seizures. These were to be annotated under Patient History, with temporal features if a new onset was mentioned, and under Seizure Frequency, when frequency was stated. In other cases as in '*She is having seizures*' the annotation was to be made under Patient History. As non-specific seizures are mentioned frequently in epilepsy clinic letters, it was decided that when seizures are mentioned in the context of seizure frequency they do not need to be annotated in Patient History, also here, only specific references to patient's seizures as '*his/her seizures*' should be annotated. This clarification helped in the annotation process and reduced the annotation time by removing the unnecessary duplication.

For Onset, When Diagnosed, and Patient History annotations, when age or time since had to be annotated, clear instructions were given for range expressions, to annotate both values, i.e. lower and upper, which is similar to the approach used by Viani et al. [225]. In their annotation of time since, based on date, however, the annotators where requested to calculate the difference themselves by using the document date, which was not the case in this project.

No other annotation guidelines seem to mention the assignment of certainty levels to the entities. As these were extracted by the algorithm it was necessary for the annotators to assign the certainty level values as attributes, the trigger phrases were provided but during the testing/revision period additional terms could be added.

The set of guidelines produced is the result of a number of annotation tests and long discussions aimed at reaching a clear understanding and agreement on the method. In general, the level of detail needed in the guidelines dependents on the annotation task, for named entity annotation without temporal relations, as when building a specialist ontology, a minimal set of instructions may be sufficient, [226] or when a single clearly defined entity such as temperature above a specific value is extracted. [227]
The list of entities, including specialist phrases for investigation results, features, and temporal relations in the epilepsy clinic letters dictated the complexity of the guidelines. In terms of scope the final document is similar to that created for CLEF project, although without the luxury of a web-based system. [224] The guidelines are equally helpful for the experts, such as clinicians and not clinically trained annotators, although each group may use them for different aspects of the annotation task. For example the list of ASM with brand and generic names helps non-clinicians, but the way dates or seizure frequency are recorded is helpful to all project participants. Apart from the prescribed structure of the annotations created they can be seen as a universal approach to annotating epilepsy clinic letters, and may require only a minimum modification.

8.1.2 Documents

In order to ensure confidentiality of the test and validation corpora all documents were pseudonymised. This involved the removal and substitution of all personal demographic information, and any terms that may potentially be identifiable, with surrogate terms before the annotation process. This approach was analogous to that used in the preparation of documents for the deidentification tasks, [228, 229] which involved the removal and substitution of protected health information (PHI), as defined by the the U.S. Health Insurance Portability and Accountability Act (HIPAA), [230] but also all details of the medical practitioners and years of dates. Although fewer categories were present in the sets used for this project, for example

no private insurance details or photographs, the task was nevertheless challenging and time consuming. Also, all names, addresses of individuals, GP practices and hospitals for the replacement were not drawn from a dataset of substitutes, but made up to look fictitious, just in case some of the documents were mistakenly thought to be real. For example locations represented real place names, but the postcodes were fully made fictitious.
This approach was essential for IDEx which is designed to extract personal information, but it was also preferred for ExECT (instead of redacting) to maintain the original structure of the documents. This process was almost totally successful, with a very small numbers of omissions involving clinicians' names.

In terms of diversity, it is very important for the clinical NLP development to be based on a wide range of document types, writing styles, content, and format to ensure the application's portability. Studies have shown that the application of clinical NLP systems within a specific domain, across different institutions is possible, [231, 232] but the lack of variation in the note type will affect its portability. [233] Documents used in this project originated from the specialist epilepsy service and contained mainly clinic letters, including those reporting investigation results. The development set contained a small number of general neurology letters and some discharge summaries, to provide variation within the set. The documents were authored by a number of clinicians with different writing styles. As they originated from a small group of neurology departments, there is a danger that they might not be fully representative of various formats used across a wider range of health boards, however, in terms of epilepsy content the language used should not be different to that of other epilepsy units.

8.1.3 Annotation tasks

All annotation tasks, apart from the IDEx validation set, were performed using the Markup application. [189] Although BRAT was tested, Markup was much more suitable for the task. The main advantage was the provision of the UMLS lookups which provided the appropriate CUIs, and the ability to use the annotated documents within GATE for the validation process. There are many other software assisted annotation tools, [234] but Markup was being developed in-house alongside ExECT and its use was mutually beneficial.

The annotation workflow for the ExECT gold standard set followed that described in other studies. [178, 235, 236] Initially, two annotation tests were performed by four annotators on sets of ten letters, followed by the measurement of the IAA. As the IAAs calculated at that stage were intended to measure the level of agreement only, without clear assessment of the annotations themselves, Fleiss' kappa was used.
Only five annotation types were considered as the sets did not contain all the categories being annotated. The first test's results, i.e. the strength of agreement, using the scale described by Landis and Koch, [183] were moderate for Diagnosis, fair for Investigations and Prescriptions, slight for Seizure Frequency, and poor for Patient History. The main reason for these disagreements was the confusion about the boundaries of categories for Patient History vs Diagnosis, lack of clarity about the scope of Patient History annotations. For example, should an abnormality given in the results of an MRI be included in Patient History, and the general difficulty in understanding how Seizure Frequency should be annotated.
The IAA following the second 10-letter annotation test was improved for all categories, apart from Diagnosis, for which only slight agreement was reached. It has to be noted that the number of annotations for this entity dropped from 31 to 20 which would also affect kappa. [237] Substantial agreement was reached for Investigations, moderate for Prescriptions and Seizure Frequency, and slight for Patient History. These results were exhaustively investigated and the nature of all errors was analysed, leading to further clarification regarding the assignment of entities to phrases, as this, on review, was the main reason for the errors. Similar experiences are reported from other projects on gold standard development. [154, 238]

Although used in the initial annotation tests, the Kappa statistic is not suitable for information retrieval, where the number of entities is unknown. [186] Instead, precision, recall, and F1 score are commonly used for such tasks. [178, 225, 239] In the final test, therefore, pair-wise F1 score was used to establish the IAA on a set

of 20 clinic letters, annotated by three trained researchers, producing two sets of results, with features (entities with all features in the correct format) and without features (an equivalent to affirmed named entity recognition (NER)), table 3.2 on page 79. For the complete annotations, the micro-averaged F1 score was 0.74, with the lowest score for specific entity of 0.57 (Onset, Patient History) and the highest score of 0.90 (When Diagnosed). For the NER the micro-averaged F1 score was 0.84, with the lowest score for a specific entity of 0.68 (Patient History) and the highest of 0.98 (Prescriptions), with Diagnosis and Investigations at 0.89 each. These were per mention scores and for some entities which have multiple mentions in each letter i.e. Patient History and Diagnosis, the IAA would be expected to be higher in a 'per letter' assessment. These results are comparable to those reached by other projects measuring the IAA for clinical text extraction. For example, medication annotation (name, type, and nine attributes) in clinical trial announcements and, as a separate task, in clinical notes, produced an overall IAA F1 score of 0.90. [178] Using FDA drug labels and clinical trial announcements, disease/disorder and sign/symptom annotations based on SNOMED CT gave F1 score for IAA of 0.86 and 0.82 for disease/disorder and sign/symptom respectively for the FDA drug labels, and 0.89 and 0.76 for the same annotation type in clinical trial notes. [178]
Named entities: disorder, examination finding, medication, and body structure, annotated in patient records in a Swedish emergency unit by two clinicians, produced F1 scores of 0.77, 0.58, 0.88, and 0.80, respectively for each category. [240]. In a study on psychosis symptoms onset, the IAA F1 score of 0.55 was reached for paragraph-level agreement, and, as onset may be mentioned many times in patient notes, 0.85 on time information agreement at the patient level. [225]. There are no studies specifically addressing IAA for epilepsy variables that can be directly compared with those achieved by the tasks reviewed here.

The final corpus of 100 clinic letters was annotated by four trained annotators, and reviewed by two annotators involved in the system development, for errors relating to format (dates, time periods, frequency), with a consensus agreement on the known 'troublesome' annotations bing reached. It was agreed that this set would represent the gold standard for the evaluation of the pipeline. During the validation process, however, a small number of errors was noticed, confirming, that despite training and very detailed guidelines, a perfect manual annotation on complex clinical text is elusive. [154] In the validation of the first ExECT pipeline a 200-letter set was used, [165] as compared to 100 documents used here, but the number of entities and features was much smaller. Also, the annotator had to highlight the entities of choice rather than assign features and allocate CUIs. Other clinical NLP tasks have used validation sets of similar size [241], however to ensure that a larger number of rare entities, such as onset or birth history, are present, a substantially bigger set would have been beneficial.

8.2 ExECT development and validation

The aim of this part of the project was the redevelopment of the ExECT pipeline to include a wider range of entities, and to synchronise the format of the features assigned. The results presented in chapter 4 are discussed here.

Apart from the entities extracted by the original ExECT pipeline, v2 produced annotations for Birth History, Onset (Epilepsy Onset), Epilepsy Cause, When Diagnosed, and Patient History, which contained major comorbidities, events that may be associated with epilepsy, and non-specific seizures. The entities were assigned more detailed features with common structure for temporal concepts that would allow for output analysis (section 4.1 on page 85) and linkage. For example, the IDEx output can be linked to the ExECT Onset output to provide Record Date and Date of Birth, so that when onset is reported as '*a year ago*' it can be converted to age.

ExECT uses lists of concepts (gazetteers) extracted from the UMLS Metathesaurus. For specific epilepsy terms it was necessary to add British English versions, variations in phrases needed for precise matching, and terms wich may represent specific concepts but were not included in the UMLS lists, such as common terms used for seizures and the latest ILAE classification terms. All of these were mapped to the appropriate UMLS CUIs. A similar process was carried out for comorbidities. Investigation results ware based on '*normal*' and '*abnormal* 'and a list of abnormalities commonly identified in EEG and MRI reports were grouped and mapped onto the abnormal result CUI for these test types. Although the lists contain a vast collection of terms relating to epilepsy they are not exhaustive and do not include all diagnostic terms that may be encountered in epilepsy clinic letters, phrasing variations, or spelling errors. This was shown in the reviews of some of the missed annotations in the validation and in the later analysis of the Epi25 cohort output. A practical solution to this is the introduction of fuzzy matching for NER including phrases identifying abnormal EEG/MRI. [242,243] Active learning (Random Forest Classifier) is used in Markup to generate suggestions of UMLS mapped ASMs with features of quantity, measurement, and frequency to assist in manual annotation. [244] A similar approach could be taken to expand the lists for the diagnostic named entities.

The pipeline was validated against the gold standard set of documents, (sections 4.2.1 on page 85 and 4.2.2 on page 87). Separate validation was performed for specific entities from Patient History i.e. febrile seizures, non-specific seizures, and selected comorbidities.
Overall results for the entire corpus (micro summary) produced precision, recall, and F1 score of 0.90, 0.86, and 0.88, respectively, with the highest F1 score (1.00) for the new variables, Birth History, Epilepsy Cause, and When Diagnosed. These results are, however, based on a very small number of annotations.
When compared to the original ExECT pipeline, bearing in mind that it extracted fewer concepts, which gave precision, recall, and F1 score of 0.91, 0.81, and 0.86 respectively, it seems that the new version's performance has only slightly improved. A more pronounced change is seen in the per letter assessment, with F1 scores for the main categories of Diagnosis, Focal seizures, Generalised seizures, and Prescription of 0.98, 0.99, 1.00, and 0.99 compared to 0.94, 0.89, 0.77, and 0.95. (The concept of 'per letter'validation is explained in section 4.2.2) In general, it is difficult to compare the summary results as the two pipelines extract different categories, or different groupings, for example, the original pipeline produced

four diagnostic entities (epilepsy, epilepsy type, focal seizure, generalised seizures) which in ExECT v2 constitute a single annotation from which specific entities such as epilepsy, seizure type, or syndrome may be derived. With Clinic Date and Date of Birth now provided by IDEx. It should also be noted that the change in classification influenced the allocation of some seizure types to specific categories of focal or generalised during the analysis.

No other studies deal with the extraction of a wide range of epilepsy variables in a single application, and comparison with other works can be made only by looking at specific aspects of epilepsy. For seizure frequency, an NLP algorithm to extract frequency expressed in quantitative values (daily, 3 per week) for an internally annotated test set, achieved precision of 0.95, recall of 0.70, and F1 score of 0.81. For an externally annotated test set, the results were lower, at 0.73, 0.22, and 0.40 for the three measures. [163] In another study, three models were used to extract seizure frequency (BERT, Bio_ClinicalBERT, and RoBERTa), expressed as a classifier 'has the patient had recent seizure?', as a quantifiable variable 'How often patient has seizures?' and, a temporal measure, 'when was the most recent seizure? '. For the classification tasks the first two models achieved 80% accuracy, and all three produced F1 score of 0.86 and 0.85, respectively for the extraction of frequency and date of last seizure. [164]

EpiDEA (Epilepsy Data Extraction System) described by Cui et al. [161] has been developed to extract epilepsy information from discharge summaries from an epilepsy monitoring unit. Evaluation for EEG patten, current ASM, and past ASM, produced F1 scores of 0.89, 0.91, and 0.89 respectively, with the overall results for precision of 0.94, recall of 0.84, and F1 score of 0.89. [161]

The extraction of epilepsy diagnosis may be compared to the extraction of diagnosis for any chronic disease, although, it has to be clear how the diagnosis is defined, in this work seizures, and not only epilepsy type or syndrome, were included in the diagnosis annotation. In a study using Health Information Text Extraction (HITEx) tool on free text EHR for asthma research, principal diagnosis

and comorbidity were extracted with the an accuracy of 82% and 87%. [155] These results are not directly comparable with F1 score, as accuracy is weighted towards true positives and true negatives.
A study using an NLP application to exclude particular diagnosis was performed using Yale cTakes extension (YTEX) and a machine-learning classifier. The aim was to identity individuals without psychogenic non-epileptic seizures (PENS) from a cohort of patients diagnosed with epilepsy. The results showed precision at 0.93, recall at 0.99, and F1 score at 0.96 for the exclusion of PENS diagnosis [245]

There has been a number of developments on automatic extraction of medication, some of them in response to the clinical NLP challenges. Uzuner et al. [246] describe the results from the i2b2 challenge for the extraction of medications, dosages, modes (routes) of administration, frequencies, durations, and reasons for administration from discharge summaries. The results are presented by category, with the best F1 scores for the top ten results, for medication from 0.80 to 0.88, for dosage from 0.80 to 0.89, for modes from 0.82 to 0.90, and frequencies from 0.81 to 0.90. Duration and reason seemed very difficult to extract, with the best F1 scores of 0.45 and 0.44 respectively, out of the top ten applications. ExECT prescription annotations were validated on a complete dose phrase i.e., medication, quantity, measurement, and frequency, which produced F1 score of 0.95 per item and 0.99 per letter, which compare really well to the best results in the i2b2 challenge.
MedEX, described by Xu et al. [139] was developed to extract medication and a range of related features. When validated on discharge summaries it produced F1 score of 0.93 for drug name, and 0.95, 0.94, and 0.96 respectively, for strength (quantity with measurement), route, and frequency. When applied to outpatient clinic visit notes (n=25), F1 scores were slightly lower, at 0.91, 0.95, 0.91, and 0.93 respectively for drug name, strength, route, and frequency. The results for duration and time were reported to be low, at 0.74 and 0.57 respectively. ASMs, apart from emergencies and to address seizure clusters, are prescribed for long term treatment, and extracting duration from a single clinic note is not that relevant. In order to measure the length of time on a particular ASM, clinic dates can be used to track prescriptions and treatment duration.

Patient history annotations contain a whole range of comorbidities and non-specific (generic) seizures i.e., those that can represent epileptic or non-epileptic seizures, such as absences, myoclonus, or simply 'seizures '. Febrile seizures are the only category annotated for negated statements (no history of febrile seizures) and for positive statements, with age when given. The validation showed 13 correctly identified cases, with F1 score of 1.00. Although based on a small number of annotations, it compares well to the best results for NER.

Results, for a selected group of comorbidities, for which there were sufficient numbers of annotations showed, apart from brain tumours, a good performance. As these entities were annotated as a single group under Patient History, it is not possible to compare these results against the benchmark from the IAA. For Patient History, as an entire category, it produced F1 score of 0.57 as compared to the validation result of 0.81.

The main group of annotations in Patient History is for non-specific seizures, which in the annotation tests reviews appeared to be quite difficult, but in the validation achieved F1 score of 0.86 per item and 0.96 per letter.

8.2.1 Onset and When diagnosed

Under the heading of Onset, annotations relating to the onset of epilepsy and specific seizures are extracted. Generic seizures and other events that could, but may not, be epileptic in nature are annotated under Patient History. If onset information is provided with these seizures, it is also extracted, so that once the outputs are produced and the context of these seizure events is established, it is possible to clarify their nature. For example, '*absences*' are annotated under Patient History, because the term is used loosely and often refers to absence-like seizures, that are not really absence seizures, however, once the context is established, a diagnosis of childhood absence epilepsy for example, they may be considered as absence seizures.

Partly for this reason, but also because there may be a significant time difference between the onset of seizures and diagnosis, When Diagnosed is extracted separately. The two annotations contain the same range of temporal features which in the output can be converted to age, i.e. time since, date (or partial date), and age.

In the validation process, Onset annotations produced a F1 score of 0.97 (n=12) and When Diagnosed of 1.00 (n=5). There are no reports of NLP algorithms that extract age of epilepsy onset or diagnosis. In order to compare the results produced here, one can only look at studies addressing onset information extraction for other diseases. For example, Mowery et al. [247] investigated the extraction of age of cancer onset from family history in familial breast and colorectal cancer. They reported a F1 score of 0.97 for combined expression of onset age (age and age range) extracted from a cohort of 28,739 records of family history. For age alone F1 score was 0.87 and for age range, 0.92. This study also reported similar issues relating to age expression that had to be addressed by this project, i.e., the conversion of common expressions of age, such as '*40s*' which have a recognised upper and lower age value, but also the challenges of less clearly defined terms, as '*post-menopausal*' or '*childhood*'. In this project such descriptive terms were converted, as much as possible, to numerical ranges and were used when compared with the SNB dataset. The same approach was taken by Viani et.al. [248] for their work on extracting temporal information to identify duration of untreated psychosis.

It has to be noted that the onset annotations extracted by ExECT relate to epilepsy and specific seizure types, hence there may be more than one onset date. This has to be remembered in the analysis of the output and it may be helpful to cross reference with When Diagnosed, as no onset age should be stated after the age of diagnosis unless it relates to specific seizure type. The ability to extract detailed age of onset, is a significant step forward for the pipeline, it provides information that is relevant to epilepsy diagnosis, treatment, and prognosis. [249–252]

8.2.2 Epilepsy cause

Epilepsy cause is annotated as a separate entity with clear triggers for causality within the rules. There were only 4 annotations extracted and F1 score was 1.00. Additional information on factors that may be linked to epilepsy can be gained from Birth History and Patient History. Birth History extracts annotations for premature birth (different degrees of prematurity) and birth injuries. F1 score for Birth History was 1.00 (n=10). Within Patient History there are annotations for head injuries, brain tumours, congenital abnormalities, and brain infections. As the requirements of the epilepsy cause rules are very strict, to avoid any false positives, it is likely that some imprecise statements of causality may be missed. For example '*symptomatic epilepsy, head injury 2010*' may be saying that the epilepsy is related to the head injury but also might not. When the diagnosis is stated as symptomatic epilepsy (general or specific) but no cause has been extracted, linking Birth History, Patient History, Onset, and Investigations may assist in clarifying the diagnosis.

8.2.3 Certainty

All diagnostic entities annotated by ExECT have, as one of the features, a level of certainty. (Section 2.4.5 provides a more in depth explanation of the concept, and B provided the list of terms used in the gold standard annotations and the ExECT v2 pipeline) These are assigned to record the certainty expressed about the entity as in '*Diagnosis: focal epilepsy, possibly temporal lobe*' which gives the diagnosis of epilepsy with the certainty level 5, and temporal lobe with the certainty level 4. Validation of certainty levels is linked to the entities they describe, especially when they are the only annotation feature i.e., no temporal relations are annotated at the same time. The results given in the validation of Diagnosis, Epilepsy Cause, Investigations, Birth History, and some annotations in Patient History (comorbidities) provide, therefore, the results for certainty evaluation.

There are limited number of studies that discuss certainty attributes for the entities extracted. Kraljevic et at. [253] report on one of the MedCAT installations that can identify concepts that are confirmed to be present or suspected, producing macro average F1 score of 0.84. Some of the concepts identified included specific epilepsy entities: '*Tonic-clonic epilepsy (disorder)*', '*Generalized epilepsy (disorder)*', '*Status epilepticus (disorder)*', '*Petit-mal epilepsy (disorder)*', '*Epilepsy (disorder)*' with their SNOMED-CT codes. The previously mentioned EpiDEA [161] adapted the NegEx algorithm to detect uncertainty which had trigger phrases of: probable, possible, likely, might have, suspected, or suspicious etc., so it is reasonable to assume that the results reported for EEG Pattern, Current ASM, and Previous ASM of F1 score at 0.89, 0.91, and 0.86 reflect the annotations of uncertainty.

8.2.4 Strengths and limitations

ExECT v2 produces a wider range of annotations than the original pipeline with a standard output for temporal relations across annotations that allows for linkage of the output data. Some of the entities extracted are not produced by any other application. The pipeline uses specialist ontologies that incorporate the latest ILAE classification and alternative spelling / phrasing which are mapped to the UMLS concepts and CUIs. The validation was performed against the gold standard set of annotations and produced results that are an improvement on the original pipeline and compare well to other systems for named entity and relations extraction.

The pipeline extracts only current medications, with the assumption that in long term all ASMs prescribed will be captured giving precise information on the dose and the length of time taken. This may not apply to individuals who have moved and/or changed their treatment centre and come to a neurology clinic with a long history of epilepsy treatment. It may be useful to capture this information as part of the patient history content.

The main limitation of this work is the small size of the gold standard corpus, which restricted the number of cases extracted under each entity, especially for the new categories. The manual construction of the gazetteers was time consuming and influenced the range of alternatives that could be created. It would be beneficial to use fuzzy matching in the process. The application was not tested on documents that originate from other health boards.

At present the algorithm is set to annotate only the positive statements, with the exception of febrile seizures. Negative statements could be added, so that when, for example, a diagnosis is actively excluded, this information is available. EEG results could contain a feature assigning the abnormal result, if stated, to '*supportive*' of a specific epilepsy type or syndrome, which would require only a small modification of the current rules. Some adjustments to the rules governing negotiations and certainty levels for long phrases may be needed to avoid errors.

8.3 IDEx

The IDEx pipeline was developed to provide a set of annotations for personal demographic information that are required to anonymise data for SAIL projects (FILE 1). It also allowed for the creation of Record Date from the extracted clinic/letter dates, which provide a timeline when the outputs from a range of annotations form sets of letters for single individuals are linked together.

The first IDEx validation was performed on a set of 200 pseudonymised clinic letters annotated in BRAT, producing F1 scores of 1.00 for Date of Birth, Hospital Number, NHS Number, and Postcode. Clinic Date and Gender were validate together with the original ExECT pipeline achieving F1 of 0.98 and 0.99 respectively. The second validation was performed against the SNB Donors database extract for 111 individuals from the Epi25 cohort. This time Hospital Number was not extracted as it was not available. The results were very similar, with F1 scores of 1.00 per individual for all entities. Finally, a random sample of 200 letters from the Epi25 cohort set was selected for manual validation of Clinic, Letter, and Record dates, with all achieving F1 scores of 1.00. The system performed very well on the clinic letters used in the validations, the question then arises how well would it perform on a wider range of documents, in different formats, such that may be produced in other health boards.

For the annotations that rely on word triggers such as NHS number, Clinic Date, or Date of Birth the algorithm should perform very well, but for the others, such as Hospital Number and Patient Postcode erroneous results may be given, as these entities very much depend on the surrounding tokens and the letter formats. For the Hospital Number rules that do not use trigger phrase, any combination of alphanumeric sequence can be annotated. For example when the application was tested on pathology reports, laboratory reference numbers were annotated in error as hospital numbers. A similar problem may affect Patient's Post Code which depends on the document structure, by being placed near patient's name and age, but this may be unique to the letter set available for this project.

The ability to extract entities without triggers, or purely based on context, is the pipeline's strength but also a weakness, which may affect its performance on documents from other centres, especially in retrieval of hospital number for which there are many patterns. The application was primarily constructed to create a dataset of personal demographic data and as such it performed very well, reaching a perfect score in a number of validations. But it depends on the letter structure and triggers being present.

In order to provide the most representative and balanced corpus of letters for the application development, documents should be acquired from different sources, ideally different health boards. The 200-letter set was mainly derived from the SBUHB, so was the development set. It was noted that hospital number annotations based of alphanumeric patterns (author's reference, punctuation, numbers) were failing when tested on letters from other organizations. Saying that the usefulness of hospital number applies to older documents and strictly only within a single institution, as one individual could have different hospital numbers in each place of treatment within the same health board, with the NHS numbers being the main patient ID.

8.4 Epi25 cohort processing – output validation and analysis

This part of the project tested the ability of the pipeline to create structured outputs that can be linked by Record Date and provide a timeline of a patient's symptoms and treatment, which is one of the aims of ExECT. As the study used a large cohort of clinic documents for the Epi25 cohort, there was also an opportunity to validate the outputs against the SNB database and to scrutinise some of the results in more detail.

8.4.1 The SNB and the Epi25 cohort

The selection of the documents was performed manually by reviewing the records for each person involved in the Epi25 study who also consented for genetic data linkage. The format of the documents was varied and required significant preprocessing, which in general was successful, with only a small number of failed splits. (Document splitting was performed when a single file containing many clinic letters was broken into separate documents). It was disappointing that for some individuals it was not possible to obtain more documents, resulting in an uneven distribution from 1 to 47 per person.

8.4.2 ExECT validation on the Epi25 SNB cohort

The validation against the SNB dataset was performed, per person, for items that are recorded as standard, such as diagnosis, seizure type, and on investigation results, as well as on some optional elements of the dataset, i.e. comorbidities. The results for the epilepsy diagnosis validation, produced precision, recall, and F1 score of 1.00, the equivalent results for focal epilepsy were 1.00, 0.92, and 0.96, and for generalised epilepsy 0.94, 1.00, and 0.97 respectively. The investigation of the errors showed that the four missing epilepsy type diagnosis could not be extracted by ExECT as the information was not available in the clinic letters available. For the two misclassified cases, one was in fact correct, i.e., the biobank database did not contain the information available in the latest clinic letter that suggested a change of diagnosis. The second misclassified diagnosis resulted in the way information relating to epilepsy type was generated for each individual, prioritising epilepsy diagnosis over seizure type. In this case the record reported IGE and complex partial seizures. It is not certain how this could be resolved, as when experimentally, the priority was switched to seizure type, there were more errors. It might be that when the filtration by date is performed first, the result would be different, but this was not tested.

Onset Having a number of letters for each individual increased the chances of capturing onset information. Although the true validation against the SNB database was not possible (section 6.2.2), the comparison of the extracted information to the SNB recorded onset produced a good results with 17 full matches and eight close matches (within the two years). Four cases from the five unmatched, were, on inspection, correctly extracted from the letters by ExECT i.e. there was a single false positive in the whole set. The discrepancy with the biobank data, results from the source and type of information recorded in the SNB database, where onset may relate to any type of seizure, specific and non-specific, derived from patient notes or reported by patients during interview, whereas ExECT strictly annotates only specific epileptic seizures, with the non-specific (generic) seizures annotated by Patient History.

There has been some research on extracting onset information, for example, cancer onset symptoms identification for familial breast and colorectal cancer reported recall of 95%. [247] Viani et.al used a rule based system for identification of psychosis onset from a corpus of mental health EHRs with disease onset mentions and reported per patient results of 71% accuracy. [225] This paper also identified the issues relating to the presence of onset information at different points of care and needing a large numbers of records for each individual to extract the information. There has been no reported research on epilepsy onset extraction. The results achieved by ExECT are very encouraging and when linked to outcomes for non-specific seizure onset provided by Patient History can supply even more information on symptoms onset that is of significant importance for correct diagnosis, treatment, and prognosis. [249, 254]

Patient History A range of entities from Patient History annotations were compared to the dataset available from the biobank. History of febrile seizure, which is recorded as standard within the biobank was compared to that captured by ExECT, in terms of sentiment analysis of the annotations rather than the number of extractions. This is because the document set available for each person did not provide the full clinical history needed to capture febrile seizure mentions. For the annotations captured the identification of correct negation status reached F1 score of 1.00, but it was based on only nine matched individuals. Although the algorithm extracted febrile seizure information for 31 people, for 22 of them the SNB database did not hold the relevant information. This might have been caused by the limited access to clinic notes during the data collection for the biobank and the fact that the clinic letters for the Epi25 cohort were collected at a later date. Although affected by small numbers, these results give some support to the previous validation outcomes for extracting febrile seizure information.

There was a similar difficulty when matching the biobank records to the information extracted by ExECT for the main comorbidities listed in the biobank: depression and dissociative seizures, which were both recorded for nine participants. Although 21 individuals with depression and 19 with dissociative seizures were identified by ExECT, only three and two, respectively, could be matched with those held in the SNB database, making validation impossible. The ExECT criteria for annotating dissociative seizures is probably stricter than that recorded in the biobank, as a significant effort was made not to capture phrases that reflected a negation of epileptic seizures such as '*some of the episodes may be non-epileptic in nature*' in favour of phrases confirming the diagnosis of dissociative seizures, as in '*she was diagnosed with non-epileptic attack disorder*'.

Investigations Although the validation of investigation results extraction against the gold standard set produced very good results, it was felt that further validation using a larger set of documents representing a greater variation of writing styles and formats would be an excellent test of the pipelines' performance. A verified collection of investigation results offered by the SNB database was a suitable choice of a dataset.

As the test results reported in clinic letters are not dated, with the date provided in the output referring to the clinic letter from which the result was extracted, not the test date, the reports had to be grouped for each individual by the test outcome of normal/abnormal. The same approach was used for the investigation records sourced from the SNB database.

The comparison was performed for the EEG and MRI results. The only outcome corresponding to the results for the 100-letters set validation was for the abnormal

EEG reports, all other were much lower. This led to a very detailed and time consuming inspection of the ExECT annotations. On reviewing all records for false negatives and false positives for each type of test and outcome it appears that the ExECT results were often correct. The reduction of the erroneous outcomes ranged from just under 50% for the false negatives for the abnormal MRI results to 100% for the false negatives for the normal EEG results.
The discrepancy between the datasets and the resulting mismatch can be explained by the selection of letters available to the study. On the one hand they did not contain the full history of investigations for many individuals, and on the other they contained some more recent documents with the latest results. It appears that the SNB database does not contain all the reports for each individual, which could be explained by the limited access to patient notes during sample collections. It may also be the case that the presence of an abnormal result supporting the diagnosis removes the necessity of recording older/later normal results.

Although this second validation did not produce the expected outcomes, it helped to identify some sources of errors. In the case of true false positives and false negatives, a number of causes were identified which can be resolved. For example, the issue of two results for different tests reported in a single phrase, which caused a number of false annotations, specific abnormalities that were missed because they were not included in the ExECT gazetteer, results that contain more than a single outcome phrase, which may need to be split into two annotations. Saying that, there are examples that may never be resolved, where an investigation report contains a degree of ambiguity, that can be clarified only by the context of previous investigations or diagnosis, or where it very much depends on the human understanding and the context, as in '*He had an MRI brain scan in April 2013 which was reassuring*'.

8.4.3 Linking ExECT outputs for longitudinal analysis

The aim of this section was to present the application of the ExECT output in investigating epilepsy at an individual and population level. Two datasets produced by ExECT were used in the analysis, Seizure Frequency and Prescriptions.

In the processing of seizure frequency all quantifiable expressions of frequency were converted to rates to construct seizure severity score (SSS). This allowed for comparison over time of the scores for different seizure types. Three individuals with a significant number of clinic appointments were then selected to illustrate the analysis and their records were reviewed. Apart from a single misinterpretation of seizure frequency all records were correct for seizure type and frequency.

Using seizure scores excludes some non-quantifiable expressions used in the clinic letters, such as frequent or infrequent and increased or decreased, which arguably suggest that the seizures still occur, i.e., the individual is not completely seizure free. To include these expressions in the analysis, one could group the output into seizure free and not seizure free records, but this would ignore seizure type and the actual frequency. In general it is difficult to compare seizure frequency across time as the way it is expressed in clinic letters is not standardised. A study by Xie et al. [164] reported on extracting text spans that contained seizure freedom, frequency expressions, and date of last seizure from clinic notes, but apart from the classification task of ‘seizure freedom yes/no’ it is not clear how the frequency information is analysed after extraction of the text spans. Similarly, Decker et al. [163] developed an algorithm for extracting seizure type and the quantitative frequency, expressed as seizures per time period, achieving precision of 0.95, recall of 0.70, and F1 score of 0.82 on the internal test. The study excluded from the outset any expressions that were not quantifiable and, from what can be seen in the lists of triggers used, did not include frequencies that are expressed as a number of seizures since a specific point in time, be it a date, point in time (e.g.,last Christmas) or last clinic. It is not clear whether the extracted frequencies were converted into a single measure.

There are no studies that address this topic and offer practical solution, and this project is novel in this respect. The SSS can be used separately for each seizure type or combined into a single measure, both were done here. For the combined scores it would be helpful to provide some adjustment for seizure type. There are measures such as Janz scale, Liverpool Seizure Severity, or the National Hospital Seizure Severity Scale (NHS3) which look at the impact of seizures on individuals but they are based on questionnaires and are not really applicable to the information held in clinic letters. [255]

The output from the Prescription annotations was processed in a similar manner, building the daily ASM dose for each drug taken and then calculating the rate of the ASM as the proportion of the recommended maintenance dose for each drug. This allowed for the ASMs to be compared but also to be mapped together with a normalised SSS values. The prescription rates were assessed for the three individuals with the seizure frequency record, and apart from a small number of errors they seemed a good reflection of the ASM taken. The errors identified related to the earlier records where less structured way of reporting prescriptions was more common, some ambiguity relating to the dose, and a clear error in the ExECT context algorithm, which will need to be address.

The ASM records for the three individuals reviewed clearly correspond to their seizure frequency. For one of the individuals it is a stunning reflection of a very difficult to control epilepsy. It would seem beneficial to link this record to the Patient history and test results' output and this would most likely be done outside of this project.

Person specific linkage allows for a clear ASM–seizure frequency record inspection that may have some application in a clinical setting. It allows for an instant review of ASM and seizure frequency by seizure type without the need for a record review.

The linkage of the seizure frequency output and ASM records for the entire cohort was made by mapping the ASM onto the SSS for all seizure types and separately for GTCS (which also included focal to bilateral convulsive seizures) and divided into monotherapy and polytherapy groups. Most individuals in the cohort were on polytherapy. Those on monotherapy seemed to experience fewer seizures apart from some high scores in a small number of records that may represent individuals at the beginning of their treatment. One of the limitations in this analysis was the uneven distribution of clinic letters and different dates of entry to the study. It might be that selecting the individuals at the same point in their treatment would have provided different results. This seems to be supported by the GTCS-ASM linkage, where the two highest scores are represented by monotherapy. In general there were even fewer records for monotherapy in the GTCS group.

Further analysis was performed for SSS-ASM for monotherapy and polytherapy by drug combinations. It is difficult to see any clear patterns in the mappings. Most of the individuals were on polytherapy and some on up to four different ASMs. This is representative of the biobank cohort as a whole, as individuals who have more difficult to control epilepsy are more likely to be included in the biobank, not only by their more frequent presence in the clinics.

This was very much an initial exploration of the linkage of the datasets produced by the ExECT pipeline. Further work will be done on groups of individuals with the same diagnosis and similar length of treatment. The most important outcome of this study is that the EXECT pipeline successfully extracted information from a large set of clinic letters relating to 111 people. Overall the process was accurate and the range of information extracted is presently not captured by other applications. The detailed exploration of seizure frequency and ASM history for a small number of individuals demonstrated the potential value of this work if applied to a larger cohort..

8.5 Genetic linkage study

This study linked genetic data, routinely collected health care data, and information extracted from clinic letters within the SAIL databank. It used the linkage to explore possible correlation between epilepsy outcomes derived from routinely collected data and the burden of rare and potentially damaging genetic variants in people with epilepsy.

8.5.1 Creating a pipeline for genetic data linkage

This part of the project aimed to create a pathway for uploading, annotation, and linking genetic data to routinely collected health care records. One of the main outcomes of the study was the successful development of a pathway for uploading

and annotating whole exome sequencing (WES) datasets within the SAIL gateway and the creation of a single table with encrypted personal identifiers which could be linked to other SAIL datasets, fig.2.10 on page 60. This was a novel development and not without some challenges relating to the secure environment of the gateway, influencing the way the variant annotation package Annovar was set up, and to the format of the data involved.
In a typical SAIL project file 1 and file 2 share a link field, 'System_ID' which is encrypted after the file is uploaded to the gateway. VCF files have a specific structure [256] and any sample identifier (Sample Alias) that could be used as a link key is held in the metadata, which is removed in the preparatory stage of annotation. The files used in the project were named with the key field, and these names had to be encrypted and then used to create a SYSTEM_ID column before being merged into a single table.
As a VCF file is essentially a text file all fields representing numeric values, such as AF or CADD should have been converted to decimal numbers when creating a DB2 table. This was not done, because the files were uploaded in an unusual way, without the specification that normally accompanies file 2 upload. This resulted in some additional scripting in the analysis when dealing with calculations and range extractions. A number of lessons were drawn from this process informing the next stage of the project, the setting up of an automated pipeline for processing of larger number of genetic variant files.

This being said, the uploading, annotation, and linkage of gene variants set the way for further work, with the possibility of selecting different indicators, reannotation with new variant datasets, and reanalysing when, in time, more clinical data became available.

8.5.2 Genetic burden, unscheduled admissions and ASM records

Out of the 111 genetic datasets uploaded, 107 were linked to GP and Hospital records, table 7.1 on page 160. The most likely reason for the missing cases is that some of the individuals in the study were registered with a GP practice that has not signed up to SAIL (some 20% of all Welsh practices do not). Additionally, two individuals had no event relating to ASM prescription in their GP records reducing the number of those linked for the therapy analysis to 105.

No significant differences, in terms of overall or gene-specific burden of rare and damaging genetic variants, between the individuals in the no admissions versus admissions and monotherapy versus polytherapy groups were found, (fig 7.3 on page 164, fig 7.4 on page 165, fig 7.5 on page 166, and fig 7.6 on page 167).

There were two genes with qualifying variants exclusively present in the unscheduled admissions and ASM polytherapy groups, *CACNA1C* and *KCNQ1*, but they were derived from fewer than 5 individuals. *CACNA1C* encodes the alpha-1-subunit of a voltage-dependent L-type calcium channel expressed in human heart and brain. Pathogenic variants in *CACNA1C* have been associated with a variety of phenotypes including cardiac rhythm disorders as well as neurodevelopmental disorders including epilepsy and epileptic encephalopathies. [257, 258] Voltage-gated calcium channels are targets for anti-seizure medications and a *CACNA1C* haplotype has been previously associated with drug resistant epilepsy in one study of Chinese Han population [259] but no other studies have showed such association in all populations. [171]
KCNQ1 encodes a voltage gated potassium channel, predominately expressed in cardiac tissue but also expressed in the brain. Pathogenic *KCNQ1* variants can cause long QT syndrome as well as epilepsy. [260, 261]

This analysis was limited by a number of factors. The study cohort was small (n=107). The length of time within the study between individuals varied, from 2 to 19 years, meaning that admission and ASM history may not be truly comparable. The study group may not be fully representative of individuals with epilepsy in Wales, with all cases having a clearly defined and confirmed diagnosis falling into the categories of non-acquired focal or generalised epilepsy, as part of the Epi25 criteria. Apart from the very complex reasons for participation [262, 263] it is likely that individuals who have more difficult to control epilepsy, and for that reason are reviewed more often in a specialist clinic, have a greater chance to donate to the biobank via clinic recruitment.

ASM therapy and unscheduled hospital admissions are approximate measures for more difficult to control epilepsy. Some people may be seizure free on polytherapy whereas some people may have multiple seizures on monotherapy. Hospital admissions may reflect other issues such as access to health care services, not just epilepsy severity.

8.5.3 Genetic burden analysis by epilepsy type

104 individuals with specific epilepsy diagnosis were linked to the genetic variants dataset. It appears that one individual's diagnosis was not recorded in the SNB database, although it was confirmed for the Epi25 submission.
The two groups were not even in size, with more than twice as many cases of non-acquired focal than generalised epilepsy. This is higher than could be expected, considering that the non-acquired focal epilepsies account for about 20% of all epilepsies, and focal epilepsies overall make up some 60% of all epilepsies. [264] With the GGE representing some 15-20% of all epilepsies in adults, it seems that the numbers in our cohort were affected by the strict selection criteria for the GGE, and by the already mentioned biobank participation (8.5.2).

The two measures used, for genetic burden, cumulative Cadd score and the number of rare and potentially damaging variants show a slightly greater burden for generalised epilepsy. Bearing in mind the very small size of the cohort, it appears that these results are in line with those showing a larger genetic burden of Ultra Rare Variants (URV)* in GEE than NAFE. [266] This result is also in line with that provided by polygenic risk score(PRS) which shows that individuals with

*URVs are those not present(1) variants not seen in the DiscovEHR database and observed only once among the combined case and control test cohort (allele count [AC] = 1) or (2) variants absent in DiscovEHR [265] and observed no more than three times in the test cohort (AC $\leq$ 3) [266] generalised epilepsy have a significantly higher burden of common risk variants associated with generalised epilepsy than patients with focal epilepsy. [267]

In the gene-based analysis for rare and potentially damaging variants in the epilepsy associated and metabolic/transporter genes there were five genes exclusively present in focal and seven in generalised epilepsy, in addition to four shared ones 7.8. There were no genes, in either of the groups, that are associated with a specific epilepsy type [47, 268]. The most common were the variants for genes from the calcium voltage-gated channel alpha1 subunit, *CACNA1C, CACNA1H, CACNA1D,* and *CACNA1C.*

8.5.4 NLP output linkage within the SAIL databank

The enrichment of routinely collected health care data with a dataset extracted from clinic letters using an NLP pipeline was the ultimate aim of this project. There were nine data tables created from the extracted ExECT annotations which after linking to the tables extracted by IDEx, and merging some of the smaller datasets into combined tables, provided four datasets. These were successfully uploaded into the SAIL databank, added to the project, and liked to GP, PEDW, and the genetic dataset page 170).

8.5.5 Seizure frequency and genetic burden

Whereas hospital admissions and ASM prescriptions are used to measures epilepsy outcomes, [269] seizure frequency provide a real picture of epilepsy severity. Seizure Severity Scores were used in the linked analysis of genetic burden and seizure frequency (section 6.5 on page 139). From the original 567 extracted annotations of seizure frequency, 429 (76%) were quantifiable and could be converted to scores. Seizure frequency reported in just under 25% of the dated records was conveyed in such a way that could not be expressed as daily/weekly rate, i.e. ‘increase’, ‘decrease’, ‘continue’. The exception to this was ‘the same’ that could be linked to a previous clinic record if the time passed between the two was not longer than six months. These unstandardised ways of reporting seizure frequency are a recognised challenge in the measurement and comparison of seizure frequency records. [164]

The last available (most recent) dated scores were produced for 100 individuals. There was no difference, in terms of cumulative CADD score and the numbers of rare and potentially damaging variants, between individuals who were seizure free for over a year and those not seizure free (having at least one seizure per year to daily seizures).

For the gene-based analysis there were a number of variants found in the not-seizure free group. For the epilepsy associated genes there was only a single gene (maximum two individuals) in the seizure free group as compared to 15 in the not seizure free group. For the metabolism/transporter genes, *ABCG2* variants were present exclusively in the not seizure free group. *ABCG2* is a ATP-binding cassette (ABC) transporter primarily associated with breast cancer [270] but has been also investigated in the context of drug resistant epilepsy, with no association being found. [270–272] A recent case control study of drug resistant epilepsy in children suggested an association with a specific variant. [273] There have not been reports of an association between the variants found in our cohort of not seizure free individuals and ASM resistance.

The way in which seizure frequency is reported in clinic letters makes it difficult to compare between individuals and for the same individual overtime. In this linkage the SSS used were based purely on frequency, without consideration being given to seizure type, although it was available in the extracted dataset. Analysis by seizure type and frequency would provide a more complete picture of seizure severity, but was not performed due to the small numbers of individuals involved. Also, the measures of seizure frequency rely on documentation during clinic visits and the numbers of clinic letters available for each individual in the cohort were not the same. It might be expected that people with more frequent seizures are reviewed more frequently in a specialist clinic, resulting in more detailed seizure frequency record.

Studies investigating genetic burden and epilepsy outcomes use different measures between ASM responders and non-responders. [170, 172] It is difficult therefore to compare the results shown here to other research. Saying that, apart from findings relating to individual variants and specific ASMs, [169] no polygenic risk for drug resistance relating to seizure response has been identified. [173]

This novel nature of this study is that it used automatically extracted seizure frequency which has been linked with the WES data and routinely collected health records. The unique aspect of this linkage is that the WES data was annotated within the secure environment of the SAIL databank.

8.6 Conclusions

Epilepsy is a common and complex disease with multiple aetiologies and significant comorbidities. Despite the developments in treatment some 30% of individuals do not become seizure free. This thesis described the development and implementation of an NLP pipeline to extract structured information from free text of epilepsy clinic letters and to make it available for research. It presented how the datasets produced could be used in clinic setting and in research, and demonstrated how they could enrich routinely collected information in a SAIL study linking genetic data to GP and hospital records to investigate potential link between genetic variants and epilepsy outcomes.

The work resulted in the creation of guidelines for the annotation of epilepsy clinic letters and producing a gold standard annotation set of 100 clinic letters.

The redevelopment of the ExECT pipeline increased the range of variables to be extracted, and created a standard output for temporal concepts, allowing for linkage and analysis of temporal expressions. Some of the new entities captures such as onset, birth history, and specific comorbidities are not extracted by any other application. The validation of the output on the gold standard set and on the selected data items from the SNB produces very good results. The errors identified were thoroughly reviewed and provided valuable material for the algorithm adjustments.

IDEx allows for extraction of personal demographic dataset to create a file 1 for the SAIL upload, and the creation of a dated record that can be used in the linkage of all output files produced by ExECT. IDEx validation on the original 200 letter set and on the extract from the Donors dataset of the SNB achieved excellent results.

In the post-processing of the ExECT outputs a set of algorithms was produced for analysing epilepsy diagnosis, prescription, and seizure frequency data, that created standard groupings and/or measures for each category, such as conversion of seizure frequency to daily frequency or seizure severity scores. These scores were used in the Epi25 dataset analysis linking seizure frequency and ASM prescription. Although no significant results could be extracted from this linkage, it provided an example of how the datasets extracted could be used in a clinic setting and in research. They allowed to observe individual seizure frequency and ASM records over time. To produce a better dataset for analysis of the output for the cohort as a whole, further work should include a selection of specific epilepsy diagnosis and similar length of treatment. The limitation to this analysis was caused by the small number of letters for some of the individuals and the variation in the length of time in the study.
It is planned to use the use the processing algorithms on a much larger dataset within the SAIL gateway on the uploaded ExECT output.

The Sail genetics project demonstrated a proof of concept by uploading, annotating, and linking VCF data to anonymised health care records, establishing a pathway that can be followed by other studies. The addition of the seizure frequency extract, which was processed inside the SAIL databank, demonstrated a successful method and the way how routinely collected data could be enriched. No genetic influence were identified in relation to admissions, ASM prescription, or seizure frequency, but the study was limited by a small sample size that can be solved with further collaboration and larger exome datasets. In further work the ExECT extracted date on comorbidities will be linked to GP dataset.

Future work

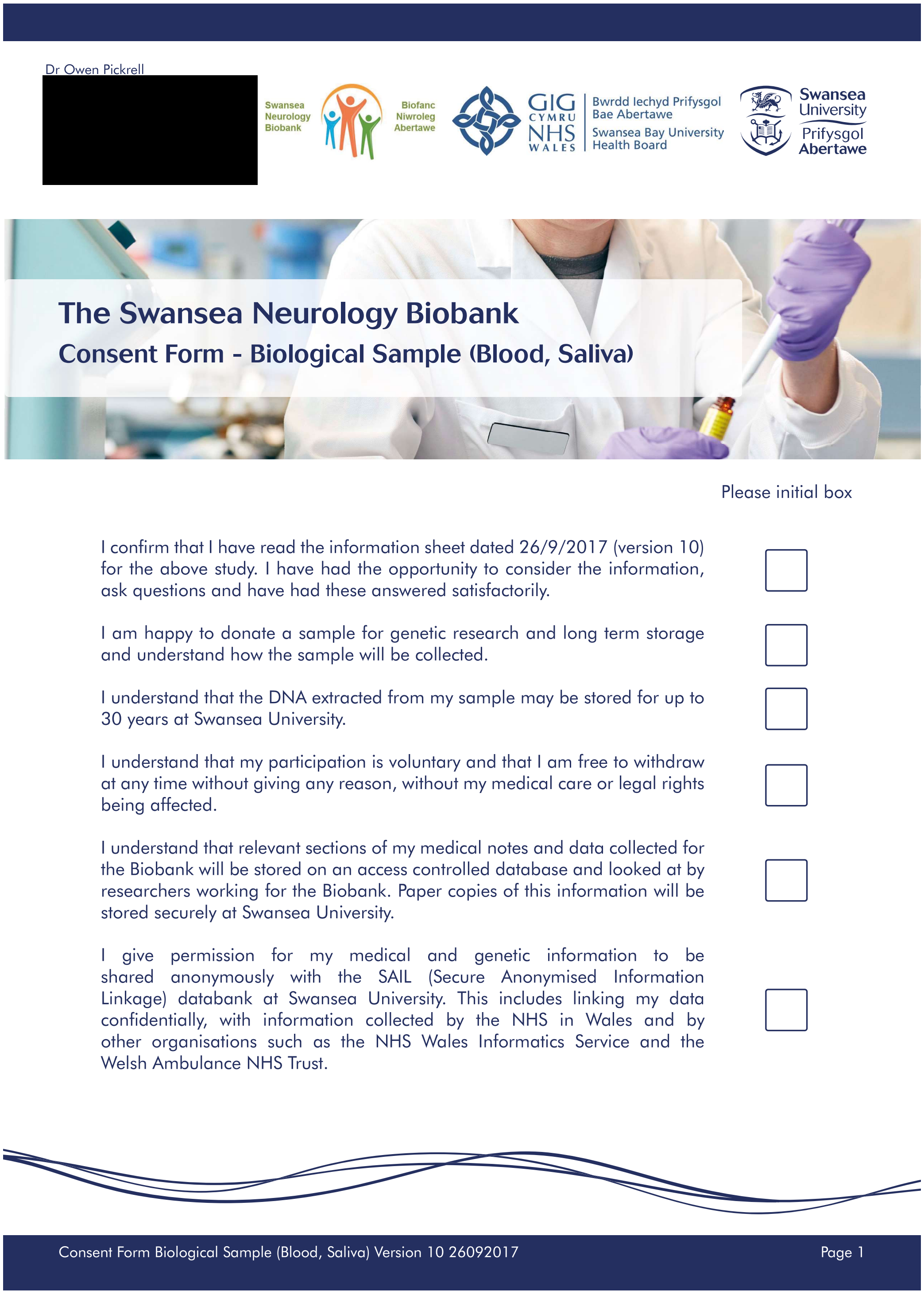
This study paves the way for further work on annotation and validation of clinic documents from other centres providing a broader range of styles and formats, which would guide future development of the pipeline. This development should consider the incorporation of the elements of machine based learning. Having access to clinic letters from across Wales, which could be processed by ExECT, would enable population level data linkage studies within SAIL. The addition of genetic data would provide an ideal scenario for research linking the NLP outputs, routinely collected health care data, and genetics, which has been explored in this project.

Appendix A

The Swansea Neurology Biobank

A.1 Consent

A.2 Biobank database



Dr Owen Pickrell

[redacted]

Swansea Neurology Biobank

Biofanc Niwroleg Abertawe

GIG CYMRU NHS WALES

Bwrdd Iechyd Prifysgol Bae Abertawe
Swansea Bay University Health Board

Swansea University
Prifysgol Abertawe

The Swansea Neurology Biobank

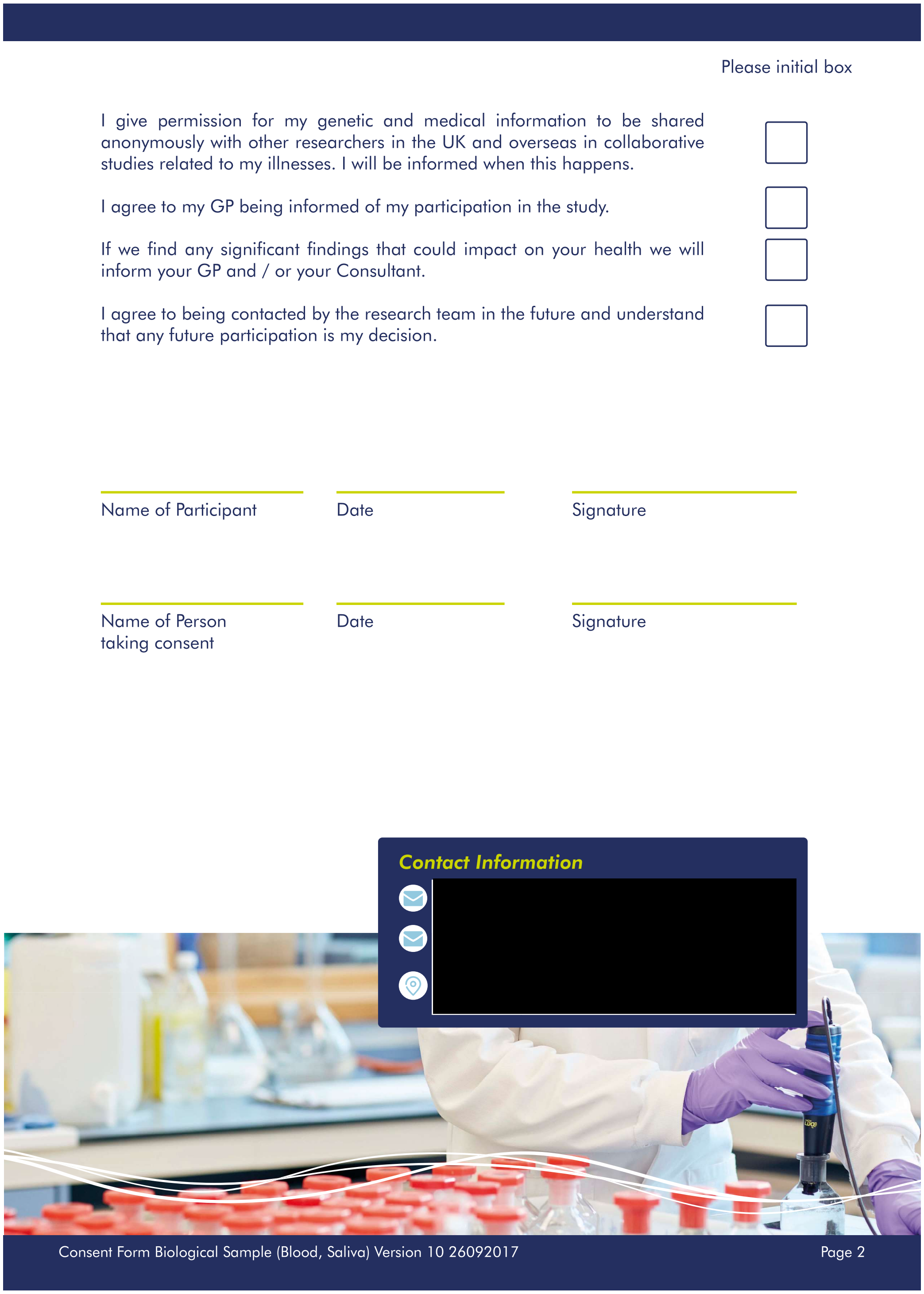
Consent Form - Biological Sample (Blood, Saliva)

Please initial box

I confirm that I have read the information sheet dated 26/9/2017 (version 10) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	□
I am happy to donate a sample for genetic research and long term storage and understand how the sample will be collected.	□
I understand that the DNA extracted from my sample may be stored for up to 30 years at Swansea University.	□
I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my medical care or legal rights being affected.	□
I understand that relevant sections of my medical notes and data collected for the Biobank will be stored on an access controlled database and looked at by researchers working for the Biobank. Paper copies of this information will be stored securely at Swansea University.	□
I give permission for my medical and genetic information to be shared anonymously with the SAIL (Secure Anonymised Information Linkage) databank at Swansea University. This includes linking my data confidentially, with information collected by the NHS in Wales and by other organisations such as the NHS Wales Informatics Service and the Welsh Ambulance NHS Trust.	□

Consent Form Biological Sample (Blood, Saliva) Version 10 26092017 Page 1

Figure A.1: Figure A1.1: SNB Consent form, page 1



Please initial box

I give permission for my genetic and medical information to be shared anonymously with other researchers in the UK and overseas in collaborative studies related to my illnesses. I will be informed when this happens. □

I agree to my GP being informed of my participation in the study. □

If we find any significant findings that could impact on your health we will inform your GP and / or your Consultant. □

I agree to being contacted by the research team in the future and understand that any future participation is my decision. □

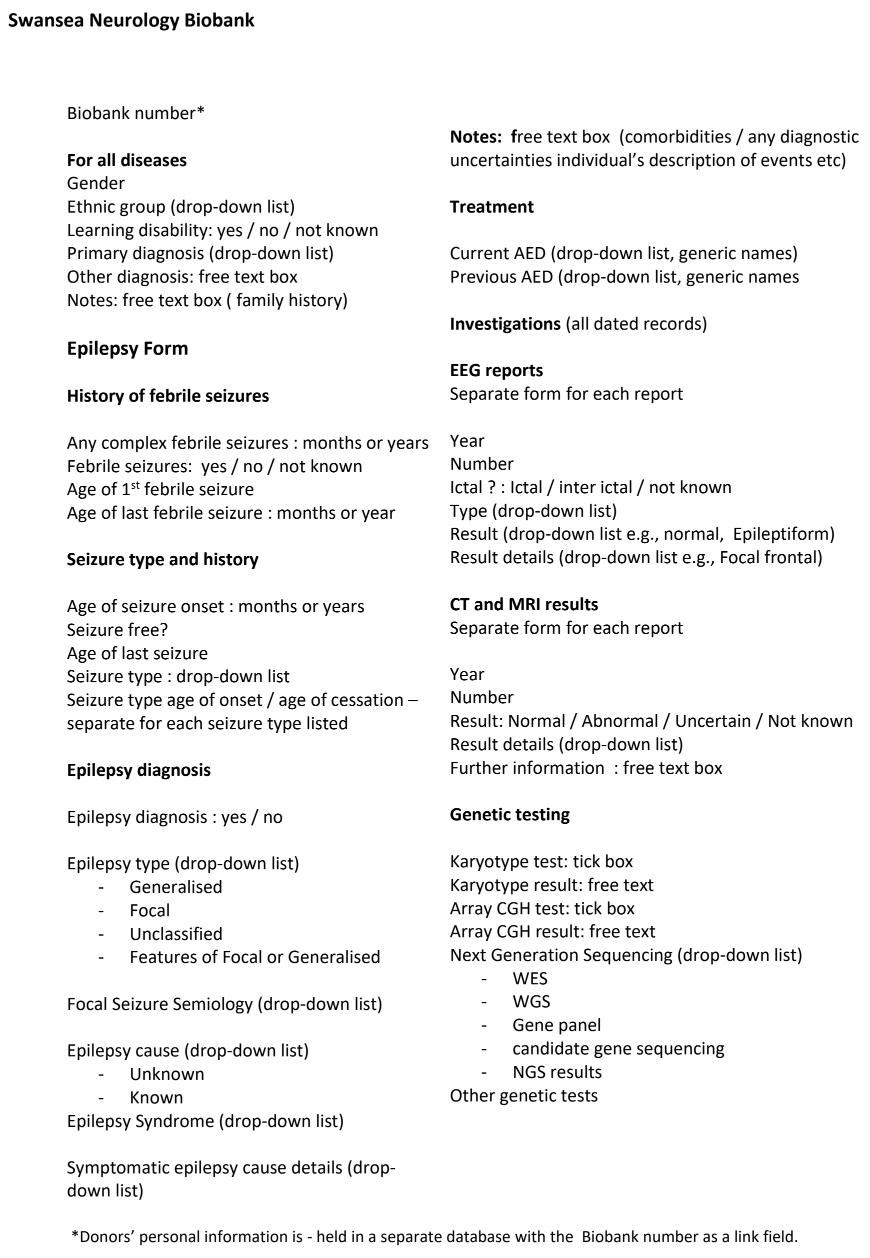
Name of Participant	Date	Signature
Name of Person taking consent	Date	Signature

Contact Information

[redacted]

Consent Form Biological Sample (Blood, Saliva) Version 10 26092017 Page 2

Figure A.2: Figure A1.1: SNB Consent form, page 2



Swansea Neurology Biobank

Biobank number*

For all diseases
Gender
Ethnic group (drop-down list)
Learning disability: yes / no / not known
Primary diagnosis (drop-down list)
Other diagnosis: free text box
Notes: free text box (family history)

Epilepsy Form

History of febrile seizures

Any complex febrile seizures : months or years
Febrile seizures: yes / no / not known
Age of 1st febrile seizure
Age of last febrile seizure : months or year

Seizure type and history

Age of seizure onset : months or years
Seizure free?
Age of last seizure
Seizure type : drop-down list
Seizure type age of onset / age of cessation – separate for each seizure type listed

Epilepsy diagnosis

Epilepsy diagnosis : yes / no

Epilepsy type (drop-down list)

- Generalised
- Focal
- Unclassified
- Features of Focal or Generalised

Focal Seizure Semiology (drop-down list)

Epilepsy cause (drop-down list)

- Unknown
- Known

Epilepsy Syndrome (drop-down list)

Symptomatic epilepsy cause details (drop-down list)

Notes: free text box (comorbidities / any diagnostic uncertainties individual's description of events etc)

Treatment

Current AED (drop-down list, generic names)
Previous AED (drop-down list, generic names

Investigations (all dated records)

EEG reports
Separate form for each report

Year
Number
Ictal ? : Ictal / inter ictal / not known
Type (drop-down list)
Result (drop-down list e.g., normal, Epileptiform)
Result details (drop-down list e.g., Focal frontal)

CT and MRI results
Separate form for each report

Year
Number
Result: Normal / Abnormal / Uncertain / Not known
Result details (drop-down list)
Further information : free text box

Genetic testing

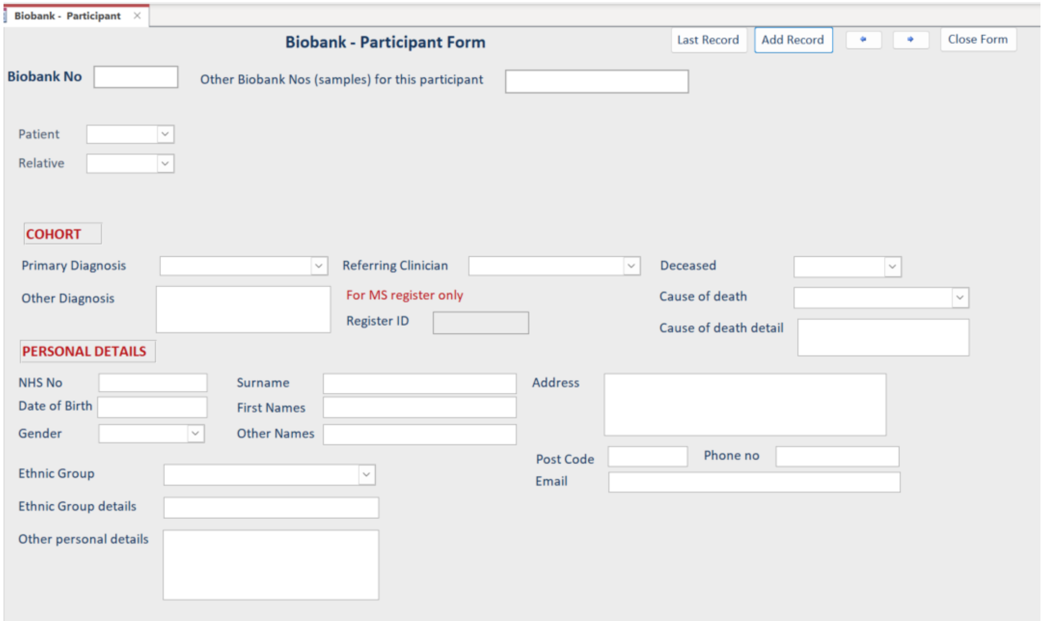
Karyotype test: tick box
Karyotype result: free text
Array CGH test: tick box
Array CGH result: free text
Next Generation Sequencing (drop-down list)

- WES
- WGS
- Gene panel
- candidate gene sequencing
- NGS results

Other genetic tests

*Donors' personal information is - held in a separate database with the Biobank number as a link field.

Figure A.3: Figure A1.2: SNB clinical database epilepsy contents



Biobank - Participant Form

Biobank No

Other Biobank Nos (samples) for this participant

Patient

Relative

COHORT

Primary Diagnosis

Referring Clinician

Deceased

Other Diagnosis

For MS register only

Register ID

Cause of death

Cause of death detail

PERSONAL DETAILS

NHS No

Date of Birth

Gender

Surname

First Names

Other Names

Address

Ethnic Group

Ethnic Group details

Other personal details

Post Code

Phone no

Email

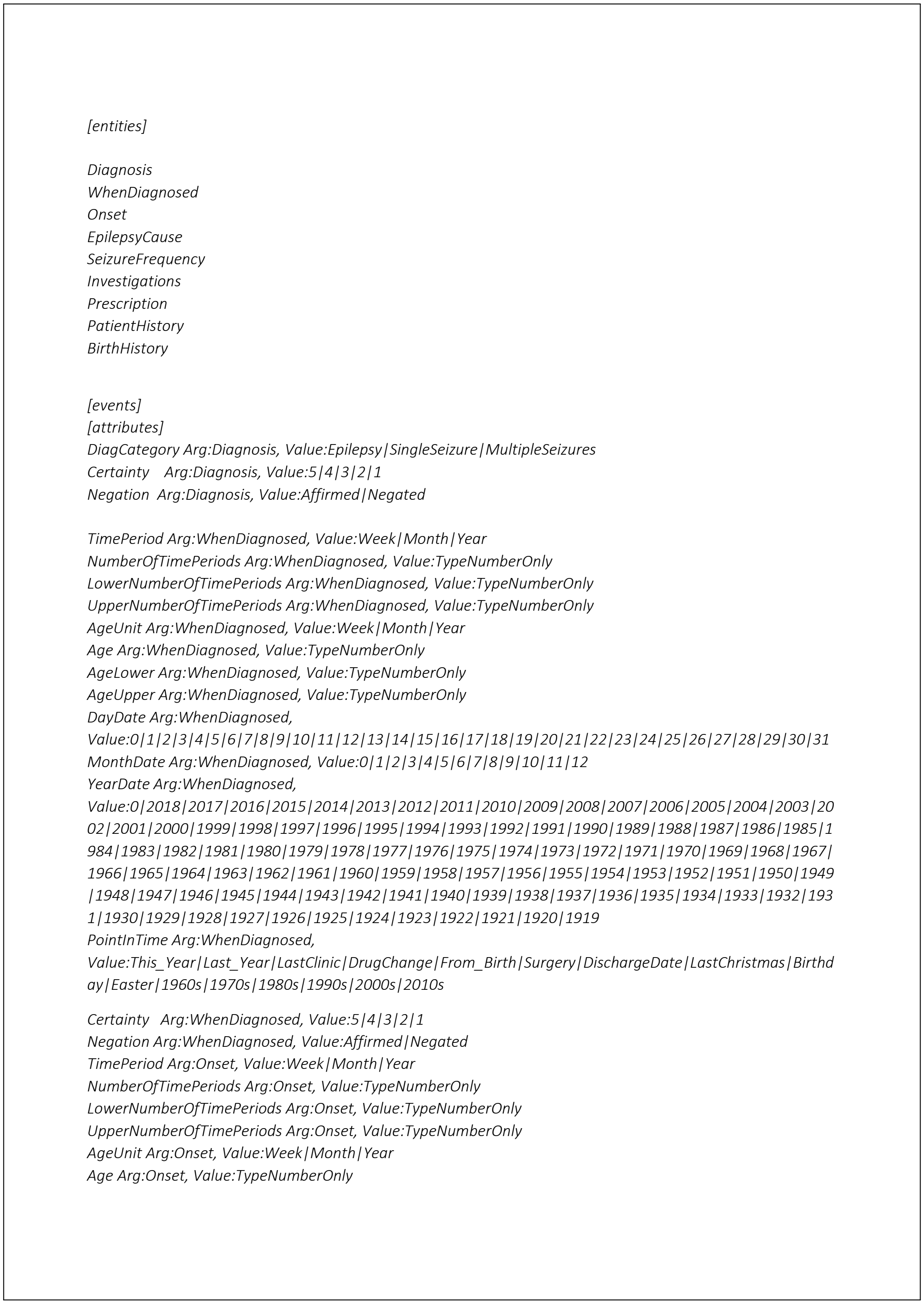
Figure A.4: Figure A1.3: SNB Donors database, Participant form

Appendix B

Gold standard annotation set

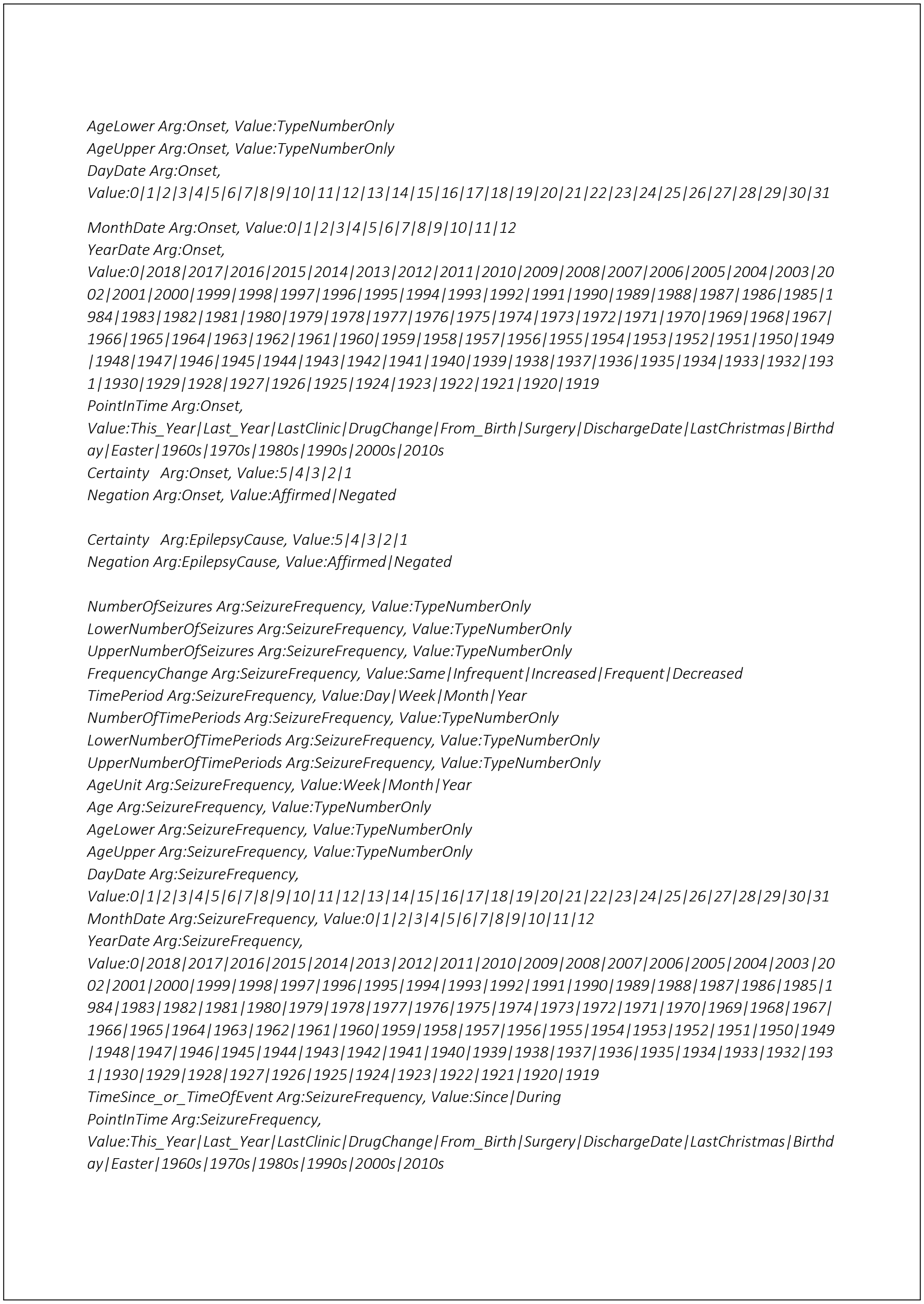
B.1 Markup configuration for epilepsy clinic letter annotation

B.2 What and how of annotating with Markup



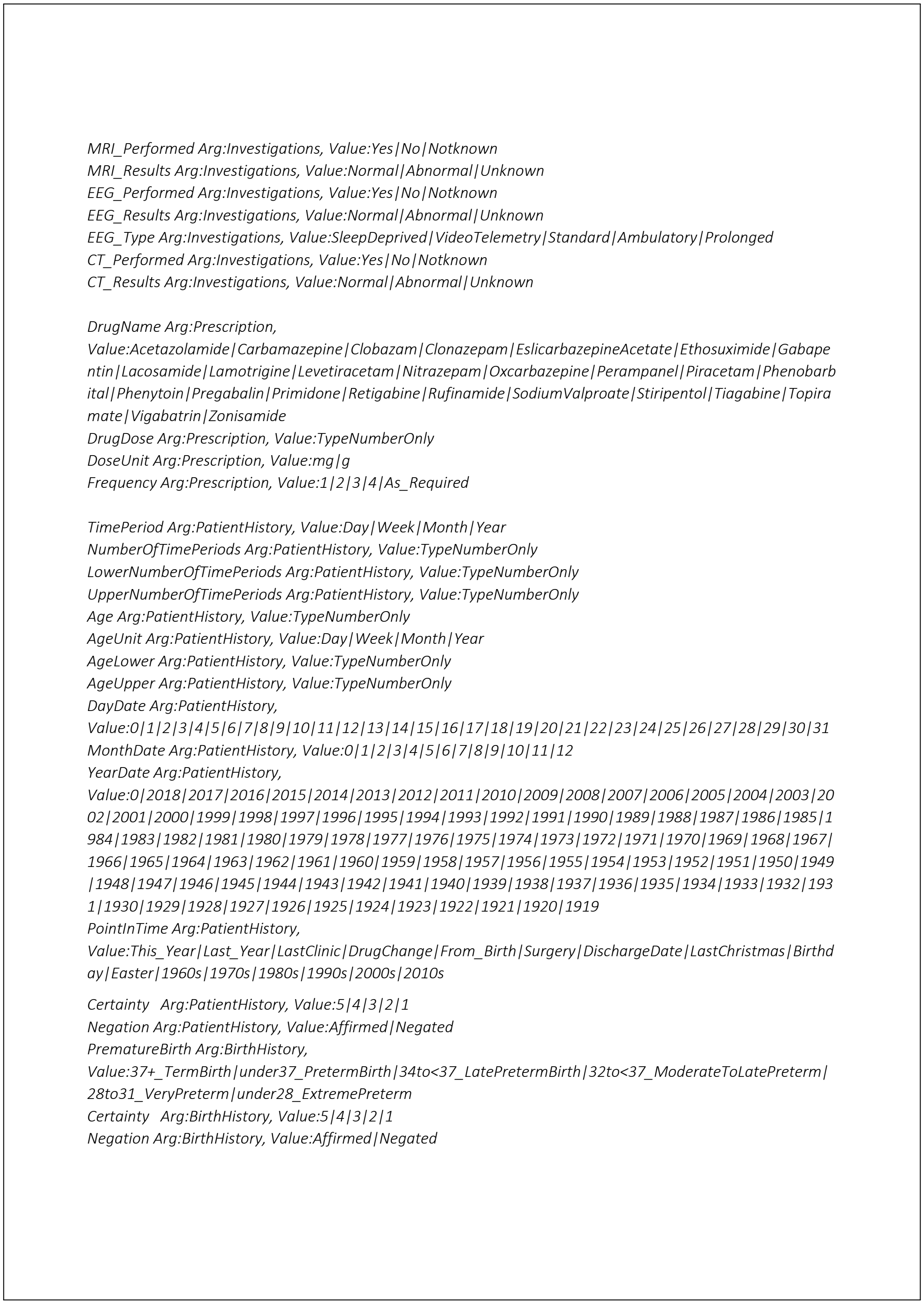
```
[entities]

Diagnosis
WhenDiagnosed
Onset
EpilepsyCause
SeizureFrequency
Investigations
Prescription
PatientHistory
BirthHistory


[events]
[attributes]
DiagCategory Arg:Diagnosis, Value:Epilepsy|SingleSeizure|MultipleSeizures
Certainty    Arg:Diagnosis, Value:5|4|3|2|1
Negation  Arg:Diagnosis, Value:Affirmed|Negated

TimePeriod Arg:WhenDiagnosed, Value:Week|Month|Year
NumberOfTimePeriods Arg:WhenDiagnosed, Value:TypeNumberOnly
LowerNumberOfTimePeriods Arg:WhenDiagnosed, Value:TypeNumberOnly
UpperNumberOfTimePeriods Arg:WhenDiagnosed, Value:TypeNumberOnly
AgeUnit Arg:WhenDiagnosed, Value:Week|Month|Year
Age Arg:WhenDiagnosed, Value:TypeNumberOnly
AgeLower Arg:WhenDiagnosed, Value:TypeNumberOnly
AgeUpper Arg:WhenDiagnosed, Value:TypeNumberOnly
DayDate Arg:WhenDiagnosed,
Value:0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31
MonthDate Arg:WhenDiagnosed, Value:0|1|2|3|4|5|6|7|8|9|10|11|12
YearDate Arg:WhenDiagnosed,
Value:0|2018|2017|2016|2015|2014|2013|2012|2011|2010|2009|2008|2007|2006|2005|2004|2003|20
02|2001|2000|1999|1998|1997|1996|1995|1994|1993|1992|1991|1990|1989|1988|1987|1986|1985|1
984|1983|1982|1981|1980|1979|1978|1977|1976|1975|1974|1973|1972|1971|1970|1969|1968|1967|
1966|1965|1964|1963|1962|1961|1960|1959|1958|1957|1956|1955|1954|1953|1952|1951|1950|1949
|1948|1947|1946|1945|1944|1943|1942|1941|1940|1939|1938|1937|1936|1935|1934|1933|1932|193
1|1930|1929|1928|1927|1926|1925|1924|1923|1922|1921|1920|1919
PointInTime Arg:WhenDiagnosed,
Value:This_Year|Last_Year|LastClinic|DrugChange|From_Birth|Surgery|DischargeDate|LastChristmas|Birthd
ay|Easter|1960s|1970s|1980s|1990s|2000s|2010s

Certainty   Arg:WhenDiagnosed, Value:5|4|3|2|1
Negation Arg:WhenDiagnosed, Value:Affirmed|Negated
TimePeriod Arg:Onset, Value:Week|Month|Year
NumberOfTimePeriods Arg:Onset, Value:TypeNumberOnly
LowerNumberOfTimePeriods Arg:Onset, Value:TypeNumberOnly
UpperNumberOfTimePeriods Arg:Onset, Value:TypeNumberOnly
AgeUnit Arg:Onset, Value:Week|Month|Year
Age Arg:Onset, Value:TypeNumberOnly
```

Figure B.1: Markup configuration page 1



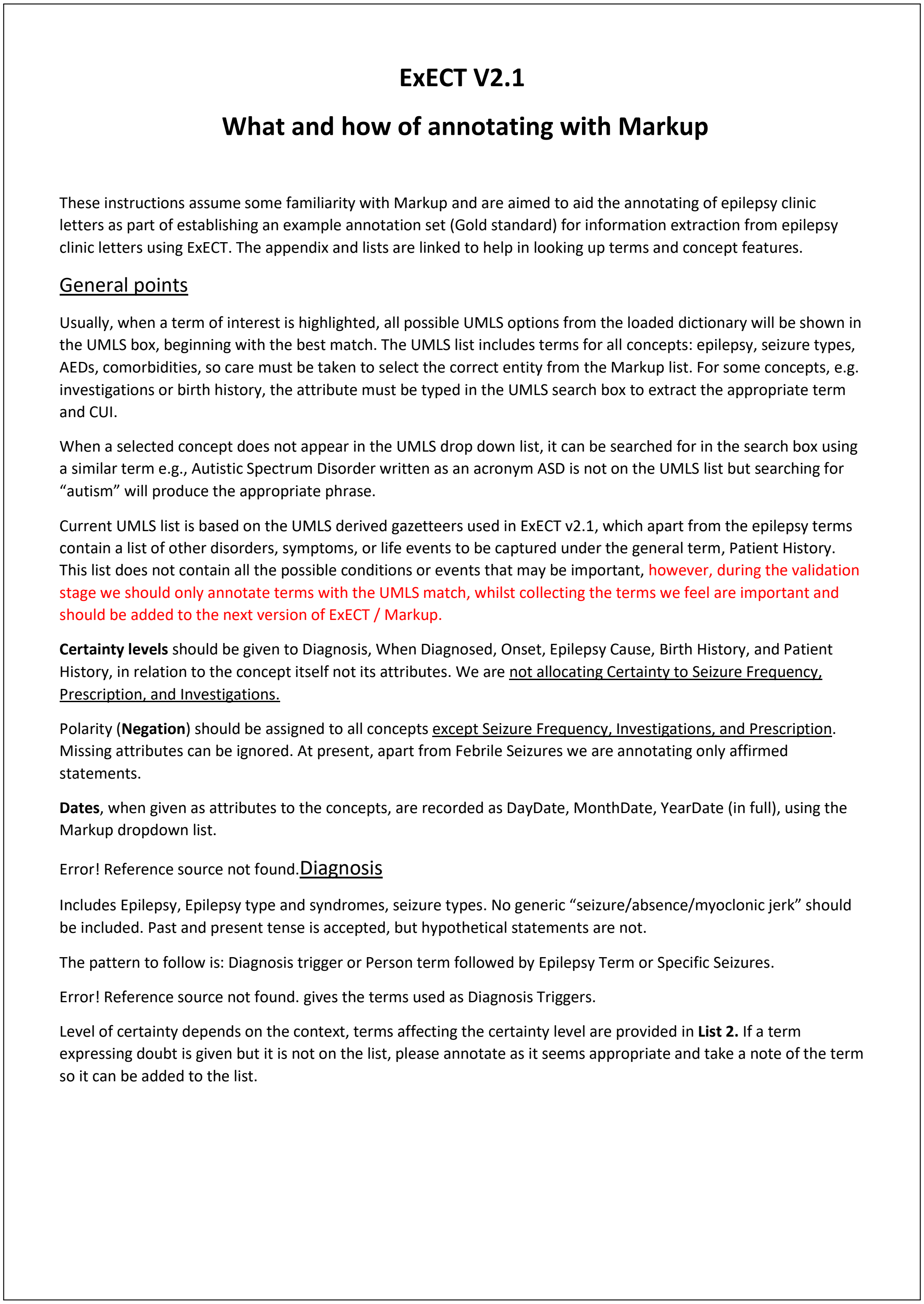
```
AgeLower Arg:Onset, Value:TypeNumberOnly
AgeUpper Arg:Onset, Value:TypeNumberOnly
DayDate Arg:Onset,
Value:0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31

MonthDate Arg:Onset, Value:0|1|2|3|4|5|6|7|8|9|10|11|12
YearDate Arg:Onset,
Value:0|2018|2017|2016|2015|2014|2013|2012|2011|2010|2009|2008|2007|2006|2005|2004|2003|20
02|2001|2000|1999|1998|1997|1996|1995|1994|1993|1992|1991|1990|1989|1988|1987|1986|1985|1
984|1983|1982|1981|1980|1979|1978|1977|1976|1975|1974|1973|1972|1971|1970|1969|1968|1967|
1966|1965|1964|1963|1962|1961|1960|1959|1958|1957|1956|1955|1954|1953|1952|1951|1950|1949
|1948|1947|1946|1945|1944|1943|1942|1941|1940|1939|1938|1937|1936|1935|1934|1933|1932|193
1|1930|1929|1928|1927|1926|1925|1924|1923|1922|1921|1920|1919
PointInTime Arg:Onset,
Value:This_Year|Last_Year|LastClinic|DrugChange|From_Birth|Surgery|DischargeDate|LastChristmas|Birthd
ay|Easter|1960s|1970s|1980s|1990s|2000s|2010s
Certainty   Arg:Onset, Value:5|4|3|2|1
Negation Arg:Onset, Value:Affirmed|Negated

Certainty   Arg:EpilepsyCause, Value:5|4|3|2|1
Negation Arg:EpilepsyCause, Value:Affirmed|Negated

NumberOfSeizures Arg:SeizureFrequency, Value:TypeNumberOnly
LowerNumberOfSeizures Arg:SeizureFrequency, Value:TypeNumberOnly
UpperNumberOfSeizures Arg:SeizureFrequency, Value:TypeNumberOnly
FrequencyChange Arg:SeizureFrequency, Value:Same|Infrequent|Increased|Frequent|Decreased
TimePeriod Arg:SeizureFrequency, Value:Day|Week|Month|Year
NumberOfTimePeriods Arg:SeizureFrequency, Value:TypeNumberOnly
LowerNumberOfTimePeriods Arg:SeizureFrequency, Value:TypeNumberOnly
UpperNumberOfTimePeriods Arg:SeizureFrequency, Value:TypeNumberOnly
AgeUnit Arg:SeizureFrequency, Value:Week|Month|Year
Age Arg:SeizureFrequency, Value:TypeNumberOnly
AgeLower Arg:SeizureFrequency, Value:TypeNumberOnly
AgeUpper Arg:SeizureFrequency, Value:TypeNumberOnly
DayDate Arg:SeizureFrequency,
Value:0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31
MonthDate Arg:SeizureFrequency, Value:0|1|2|3|4|5|6|7|8|9|10|11|12
YearDate Arg:SeizureFrequency,
Value:0|2018|2017|2016|2015|2014|2013|2012|2011|2010|2009|2008|2007|2006|2005|2004|2003|20
02|2001|2000|1999|1998|1997|1996|1995|1994|1993|1992|1991|1990|1989|1988|1987|1986|1985|1
984|1983|1982|1981|1980|1979|1978|1977|1976|1975|1974|1973|1972|1971|1970|1969|1968|1967|
1966|1965|1964|1963|1962|1961|1960|1959|1958|1957|1956|1955|1954|1953|1952|1951|1950|1949
|1948|1947|1946|1945|1944|1943|1942|1941|1940|1939|1938|1937|1936|1935|1934|1933|1932|193
1|1930|1929|1928|1927|1926|1925|1924|1923|1922|1921|1920|1919
TimeSince_or_TimeOfEvent Arg:SeizureFrequency, Value:Since|During
PointInTime Arg:SeizureFrequency,
Value:This_Year|Last_Year|LastClinic|DrugChange|From_Birth|Surgery|DischargeDate|LastChristmas|Birthd
ay|Easter|1960s|1970s|1980s|1990s|2000s|2010s
```

Figure B.1: Markup configuration page 2



```
MRI_Performed Arg:Investigations, Value:Yes|No|Notknown
MRI_Results Arg:Investigations, Value:Normal|Abnormal|Unknown
EEG_Performed Arg:Investigations, Value:Yes|No|Notknown
EEG_Results Arg:Investigations, Value:Normal|Abnormal|Unknown
EEG_Type Arg:Investigations, Value:SleepDeprived|VideoTelemetry|Standard|Ambulatory|Prolonged
CT_Performed Arg:Investigations, Value:Yes|No|Notknown
CT_Results Arg:Investigations, Value:Normal|Abnormal|Unknown

DrugName Arg:Prescription,
Value:Acetazolamide|Carbamazepine|Clobazam|Clonazepam|EslicarbazepineAcetate|Ethosuximide|Gabapentin|Lacosamide|Lamotrigine|Levetiracetam|Nitrazepam|Oxcarbazepine|Perampanel|Piracetam|Phenobarbital|Phenytoin|Pregabalin|Primidone|Retigabine|Rufinamide|SodiumValproate|Stiripentol|Tiagabine|Topiramate|Vigabatrin|Zonisamide
DrugDose Arg:Prescription, Value:TypeNumberOnly
DoseUnit Arg:Prescription, Value:mg|g
Frequency Arg:Prescription, Value:1|2|3|4|As_Required

TimePeriod Arg:PatientHistory, Value:Day|Week|Month|Year
NumberOfTimePeriods Arg:PatientHistory, Value:TypeNumberOnly
LowerNumberOfTimePeriods Arg:PatientHistory, Value:TypeNumberOnly
UpperNumberOfTimePeriods Arg:PatientHistory, Value:TypeNumberOnly
Age Arg:PatientHistory, Value:TypeNumberOnly
AgeUnit Arg:PatientHistory, Value:Day|Week|Month|Year
AgeLower Arg:PatientHistory, Value:TypeNumberOnly
AgeUpper Arg:PatientHistory, Value:TypeNumberOnly
DayDate Arg:PatientHistory,
Value:0|1|2|3|4|5|6|7|8|9|10|11|12|13|14|15|16|17|18|19|20|21|22|23|24|25|26|27|28|29|30|31
MonthDate Arg:PatientHistory, Value:0|1|2|3|4|5|6|7|8|9|10|11|12
YearDate Arg:PatientHistory,
Value:0|2018|2017|2016|2015|2014|2013|2012|2011|2010|2009|2008|2007|2006|2005|2004|2003|2002|2001|2000|1999|1998|1997|1996|1995|1994|1993|1992|1991|1990|1989|1988|1987|1986|1985|1984|1983|1982|1981|1980|1979|1978|1977|1976|1975|1974|1973|1972|1971|1970|1969|1968|1967|1966|1965|1964|1963|1962|1961|1960|1959|1958|1957|1956|1955|1954|1953|1952|1951|1950|1949|1948|1947|1946|1945|1944|1943|1942|1941|1940|1939|1938|1937|1936|1935|1934|1933|1932|1931|1930|1929|1928|1927|1926|1925|1924|1923|1922|1921|1920|1919
PointInTime Arg:PatientHistory,
Value:This_Year|Last_Year|LastClinic|DrugChange|From_Birth|Surgery|DischargeDate|LastChristmas|Birthday|Easter|1960s|1970s|1980s|1990s|2000s|2010s
Certainty   Arg:PatientHistory, Value:5|4|3|2|1
Negation Arg:PatientHistory, Value:Affirmed|Negated
PrematureBirth Arg:BirthHistory,
Value:37+_TermBirth|under37_PretermBirth|34to<37_LatePretermBirth|32to<37_ModerateToLatePreterm|28to31_VeryPreterm|under28_ExtremePreterm
Certainty   Arg:BirthHistory, Value:5|4|3|2|1
Negation Arg:BirthHistory, Value:Affirmed|Negated
```

Figure B.1: Markup configuration page 3



ExECT V2.1

What and how of annotating with Markup

These instructions assume some familiarity with Markup and are aimed to aid the annotating of epilepsy clinic letters as part of establishing an example annotation set (Gold standard) for information extraction from epilepsy clinic letters using ExECT. The appendix and lists are linked to help in looking up terms and concept features.

General points

Usually, when a term of interest is highlighted, all possible UMLS options from the loaded dictionary will be shown in the UMLS box, beginning with the best match. The UMLS list includes terms for all concepts: epilepsy, seizure types, AEDs, comorbidities, so care must be taken to select the correct entity from the Markup list. For some concepts, e.g. investigations or birth history, the attribute must be typed in the UMLS search box to extract the appropriate term and CUI.

When a selected concept does not appear in the UMLS drop down list, it can be searched for in the search box using a similar term e.g., Autistic Spectrum Disorder written as an acronym ASD is not on the UMLS list but searching for “autism” will produce the appropriate phrase.

Current UMLS list is based on the UMLS derived gazetteers used in ExECT v2.1, which apart from the epilepsy terms contain a list of other disorders, symptoms, or life events to be captured under the general term, Patient History. This list does not contain all the possible conditions or events that may be important, however, during the validation stage we should only annotate terms with the UMLS match, whilst collecting the terms we feel are important and should be added to the next version of ExECT / Markup.

Certainty levels should be given to Diagnosis, When Diagnosed, Onset, Epilepsy Cause, Birth History, and Patient History, in relation to the concept itself not its attributes. We are not allocating Certainty to Seizure Frequency, Prescription, and Investigations.

Polarity (**Negation**) should be assigned to all concepts except Seizure Frequency, Investigations, and Prescription. Missing attributes can be ignored. At present, apart from Febrile Seizures we are annotating only affirmed statements.

Dates, when given as attributes to the concepts, are recorded as DayDate, MonthDate, YearDate (in full), using the Markup dropdown list.

Error! Reference source not found.Diagnosis

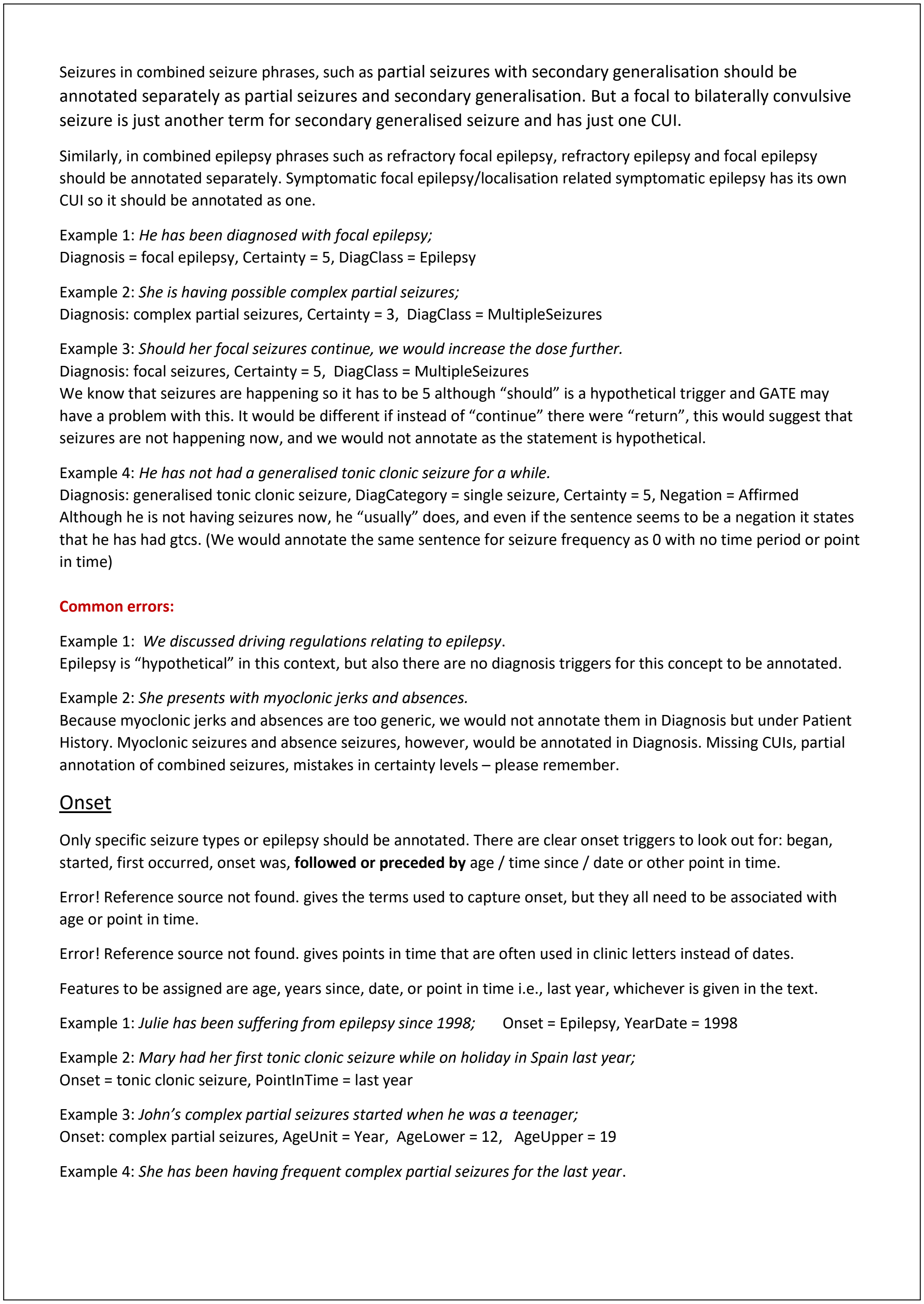
Includes Epilepsy, Epilepsy type and syndromes, seizure types. No generic “seizure/absence/myoclonic jerk” should be included. Past and present tense is accepted, but hypothetical statements are not.

The pattern to follow is: Diagnosis trigger or Person term followed by Epilepsy Term or Specific Seizures.

Error! Reference source not found. gives the terms used as Diagnosis Triggers.

Level of certainty depends on the context, terms affecting the certainty level are provided in **List 2.** If a term expressing doubt is given but it is not on the list, please annotate as it seems appropriate and take a note of the term so it can be added to the list.

Figure B.2: Annotation guidelines page 1



Seizures in combined seizure phrases, such as partial seizures with secondary generalisation should be annotated separately as partial seizures and secondary generalisation. But a focal to bilaterally convulsive seizure is just another term for secondary generalised seizure and has just one CUI.

Similarly, in combined epilepsy phrases such as refractory focal epilepsy, refractory epilepsy and focal epilepsy should be annotated separately. Symptomatic focal epilepsy/localisation related symptomatic epilepsy has its own CUI so it should be annotated as one.

Example 1: *He has been diagnosed with focal epilepsy;*
Diagnosis = focal epilepsy, Certainty = 5, DiagClass = Epilepsy

Example 2: *She is having possible complex partial seizures;*
Diagnosis: complex partial seizures, Certainty = 3, DiagClass = MultipleSeizures

Example 3: *Should her focal seizures continue, we would increase the dose further.*
Diagnosis: focal seizures, Certainty = 5, DiagClass = MultipleSeizures
We know that seizures are happening so it has to be 5 although "should" is a hypothetical trigger and GATE may have a problem with this. It would be different if instead of "continue" there were "return", this would suggest that seizures are not happening now, and we would not annotate as the statement is hypothetical.

Example 4: *He has not had a generalised tonic clonic seizure for a while.*
Diagnosis: generalised tonic clonic seizure, DiagCategory = single seizure, Certainty = 5, Negation = Affirmed
Although he is not having seizures now, he "usually" does, and even if the sentence seems to be a negation it states that he has had gtcs. (We would annotate the same sentence for seizure frequency as 0 with no time period or point in time)

Common errors:

Example 1: *We discussed driving regulations relating to epilepsy*.
Epilepsy is "hypothetical" in this context, but also there are no diagnosis triggers for this concept to be annotated.

Example 2: *She presents with myoclonic jerks and absences.*
Because myoclonic jerks and absences are too generic, we would not annotate them in Diagnosis but under Patient History. Myoclonic seizures and absence seizures, however, would be annotated in Diagnosis. Missing CUIs, partial annotation of combined seizures, mistakes in certainty levels – please remember.

Onset

Only specific seizure types or epilepsy should be annotated. There are clear onset triggers to look out for: began, started, first occurred, onset was, **followed or preceded by** age / time since / date or other point in time.

Error! Reference source not found. gives the terms used to capture onset, but they all need to be associated with age or point in time.

Error! Reference source not found. gives points in time that are often used in clinic letters instead of dates.

Features to be assigned are age, years since, date, or point in time i.e., last year, whichever is given in the text.

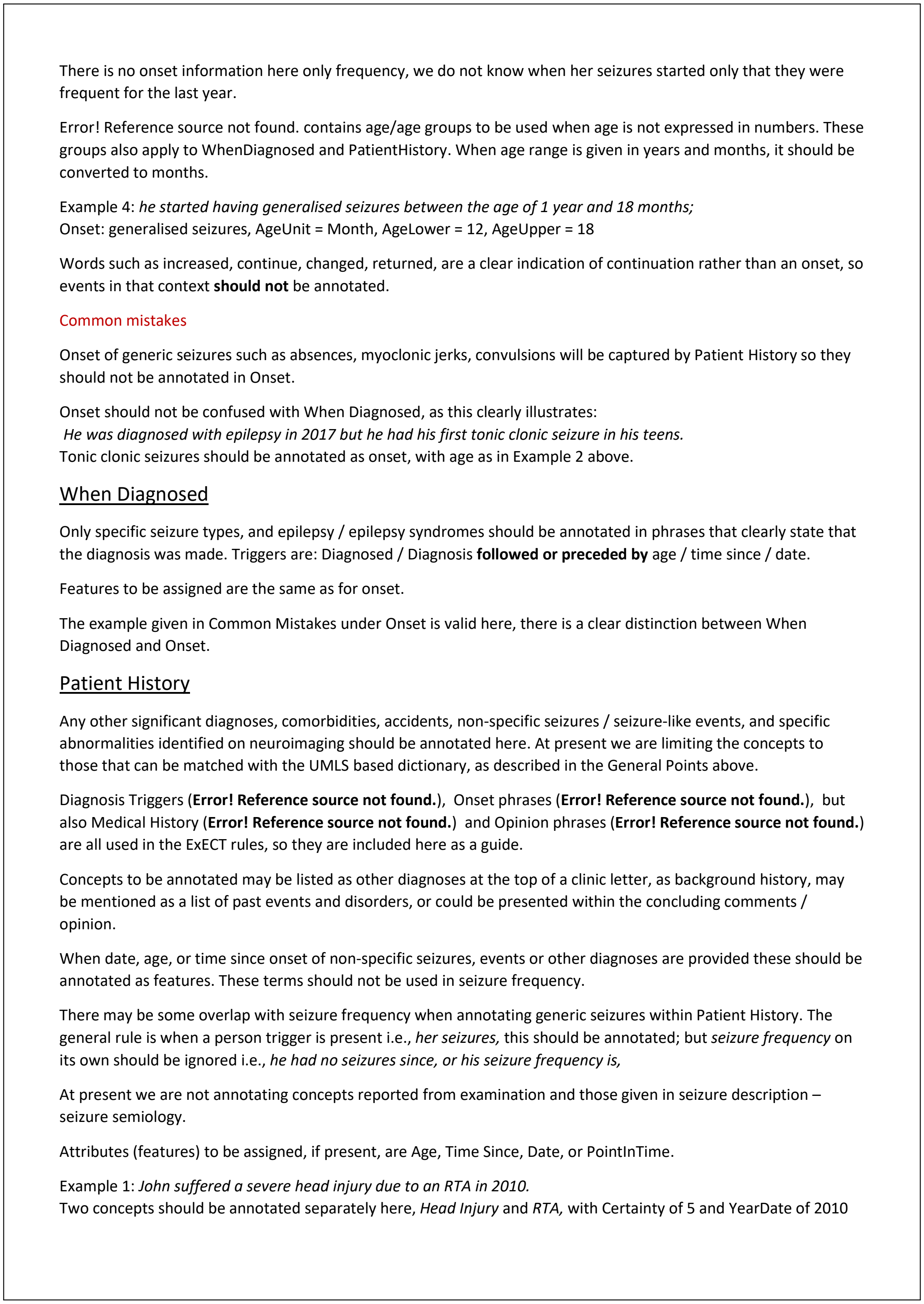
Example 1: *Julie has been suffering from epilepsy since 1998;* Onset = Epilepsy, YearDate = 1998

Example 2: *Mary had her first tonic clonic seizure while on holiday in Spain last year;*
Onset = tonic clonic seizure, PointInTime = last year

Example 3: *John's complex partial seizures started when he was a teenager;*
Onset: complex partial seizures, AgeUnit = Year, AgeLower = 12, AgeUpper = 19

Example 4: *She has been having frequent complex partial seizures for the last year*.

Figure B.2: Annotation guidelines page 2



There is no onset information here only frequency, we do not know when her seizures started only that they were frequent for the last year.

Error! Reference source not found. contains age/age groups to be used when age is not expressed in numbers. These groups also apply to WhenDiagnosed and PatientHistory. When age range is given in years and months, it should be converted to months.

Example 4: *he started having generalised seizures between the age of 1 year and 18 months;*
Onset: generalised seizures, AgeUnit = Month, AgeLower = 12, AgeUpper = 18

Words such as increased, continue, changed, returned, are a clear indication of continuation rather than an onset, so events in that context **should not** be annotated.

Common mistakes

Onset of generic seizures such as absences, myoclonic jerks, convulsions will be captured by Patient History so they should not be annotated in Onset.

Onset should not be confused with When Diagnosed, as this clearly illustrates:
He was diagnosed with epilepsy in 2017 but he had his first tonic clonic seizure in his teens.
Tonic clonic seizures should be annotated as onset, with age as in Example 2 above.

When Diagnosed

Only specific seizure types, and epilepsy / epilepsy syndromes should be annotated in phrases that clearly state that the diagnosis was made. Triggers are: Diagnosed / Diagnosis **followed or preceded by** age / time since / date.

Features to be assigned are the same as for onset.

The example given in Common Mistakes under Onset is valid here, there is a clear distinction between When Diagnosed and Onset.

Patient History

Any other significant diagnoses, comorbidities, accidents, non-specific seizures / seizure-like events, and specific abnormalities identified on neuroimaging should be annotated here. At present we are limiting the concepts to those that can be matched with the UMLS based dictionary, as described in the General Points above.

Diagnosis Triggers (**Error! Reference source not found.**), Onset phrases (**Error! Reference source not found.**), but also Medical History (**Error! Reference source not found.**) and Opinion phrases (**Error! Reference source not found.**) are all used in the ExECT rules, so they are included here as a guide.

Concepts to be annotated may be listed as other diagnoses at the top of a clinic letter, as background history, may be mentioned as a list of past events and disorders, or could be presented within the concluding comments / opinion.

When date, age, or time since onset of non-specific seizures, events or other diagnoses are provided these should be annotated as features. These terms should not be used in seizure frequency.

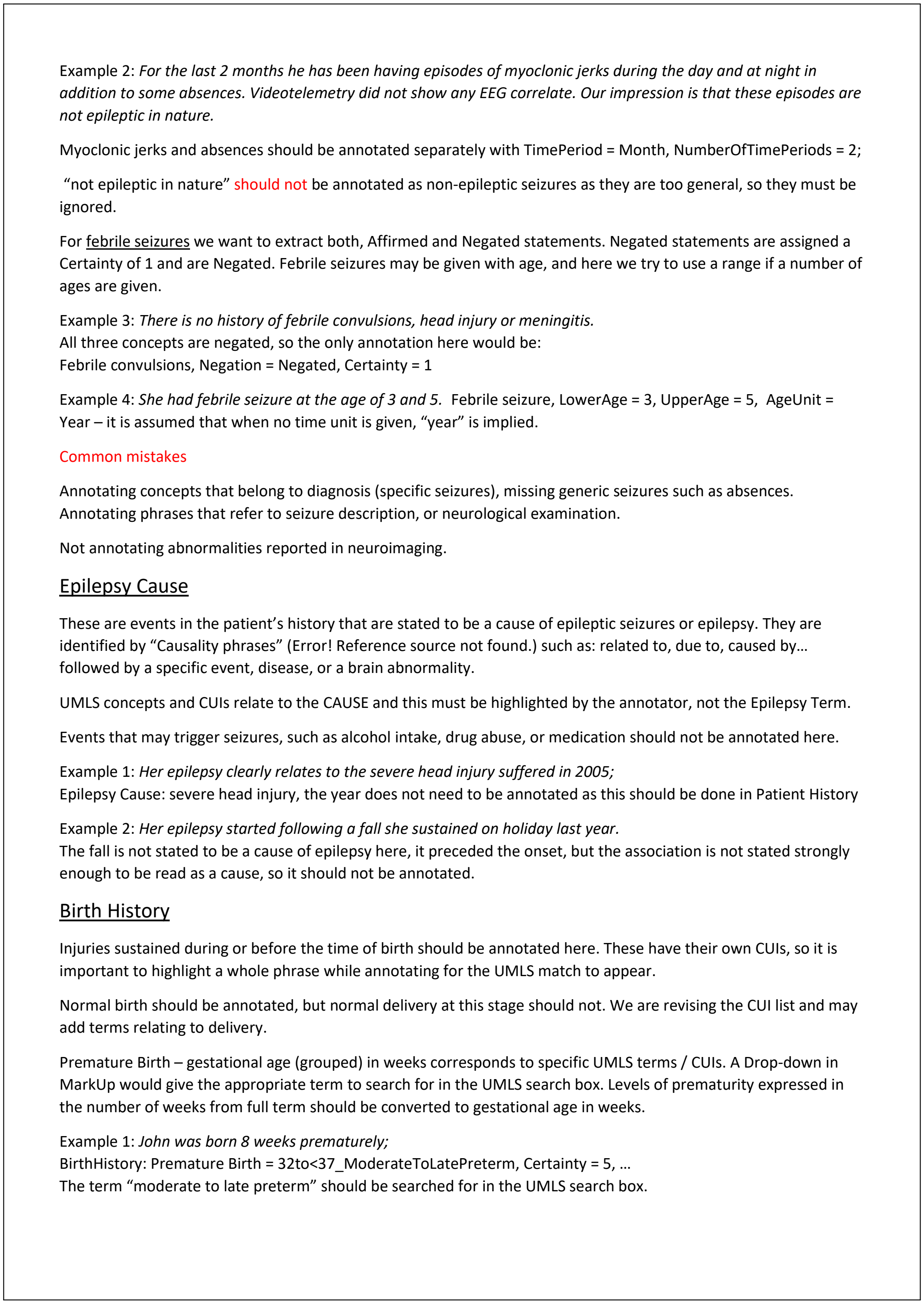
There may be some overlap with seizure frequency when annotating generic seizures within Patient History. The general rule is when a person trigger is present i.e., *her seizures,* this should be annotated; but *seizure frequency* on its own should be ignored i.e., *he had no seizures since, or his seizure frequency is,*

At present we are not annotating concepts reported from examination and those given in seizure description – seizure semiology.

Attributes (features) to be assigned, if present, are Age, Time Since, Date, or PointInTime.

Example 1: *John suffered a severe head injury due to an RTA in 2010.*
Two concepts should be annotated separately here, *Head Injury* and *RTA,* with Certainty of 5 and YearDate of 2010

Figure B.2: Annotation guidelines page 3



Example 2: *For the last 2 months he has been having episodes of myoclonic jerks during the day and at night in addition to some absences. Videotelemetry did not show any EEG correlate. Our impression is that these episodes are not epileptic in nature.*

Myoclonic jerks and absences should be annotated separately with TimePeriod = Month, NumberOfTimePeriods = 2;

"not epileptic in nature" should not be annotated as non-epileptic seizures as they are too general, so they must be ignored.

For febrile seizures we want to extract both, Affirmed and Negated statements. Negated statements are assigned a Certainty of 1 and are Negated. Febrile seizures may be given with age, and here we try to use a range if a number of ages are given.

Example 3: *There is no history of febrile convulsions, head injury or meningitis.*
All three concepts are negated, so the only annotation here would be:
Febrile convulsions, Negation = Negated, Certainty = 1

Example 4: *She had febrile seizure at the age of 3 and 5.* Febrile seizure, LowerAge = 3, UpperAge = 5, AgeUnit = Year – it is assumed that when no time unit is given, "year" is implied.

Common mistakes

Annotating concepts that belong to diagnosis (specific seizures), missing generic seizures such as absences. Annotating phrases that refer to seizure description, or neurological examination.

Not annotating abnormalities reported in neuroimaging.

Epilepsy Cause

These are events in the patient's history that are stated to be a cause of epileptic seizures or epilepsy. They are identified by "Causality phrases" (Error! Reference source not found.) such as: related to, due to, caused by… followed by a specific event, disease, or a brain abnormality.

UMLS concepts and CUIs relate to the CAUSE and this must be highlighted by the annotator, not the Epilepsy Term.

Events that may trigger seizures, such as alcohol intake, drug abuse, or medication should not be annotated here.

Example 1: *Her epilepsy clearly relates to the severe head injury suffered in 2005;*
Epilepsy Cause: severe head injury, the year does not need to be annotated as this should be done in Patient History

Example 2: *Her epilepsy started following a fall she sustained on holiday last year.*
The fall is not stated to be a cause of epilepsy here, it preceded the onset, but the association is not stated strongly enough to be read as a cause, so it should not be annotated.

Birth History

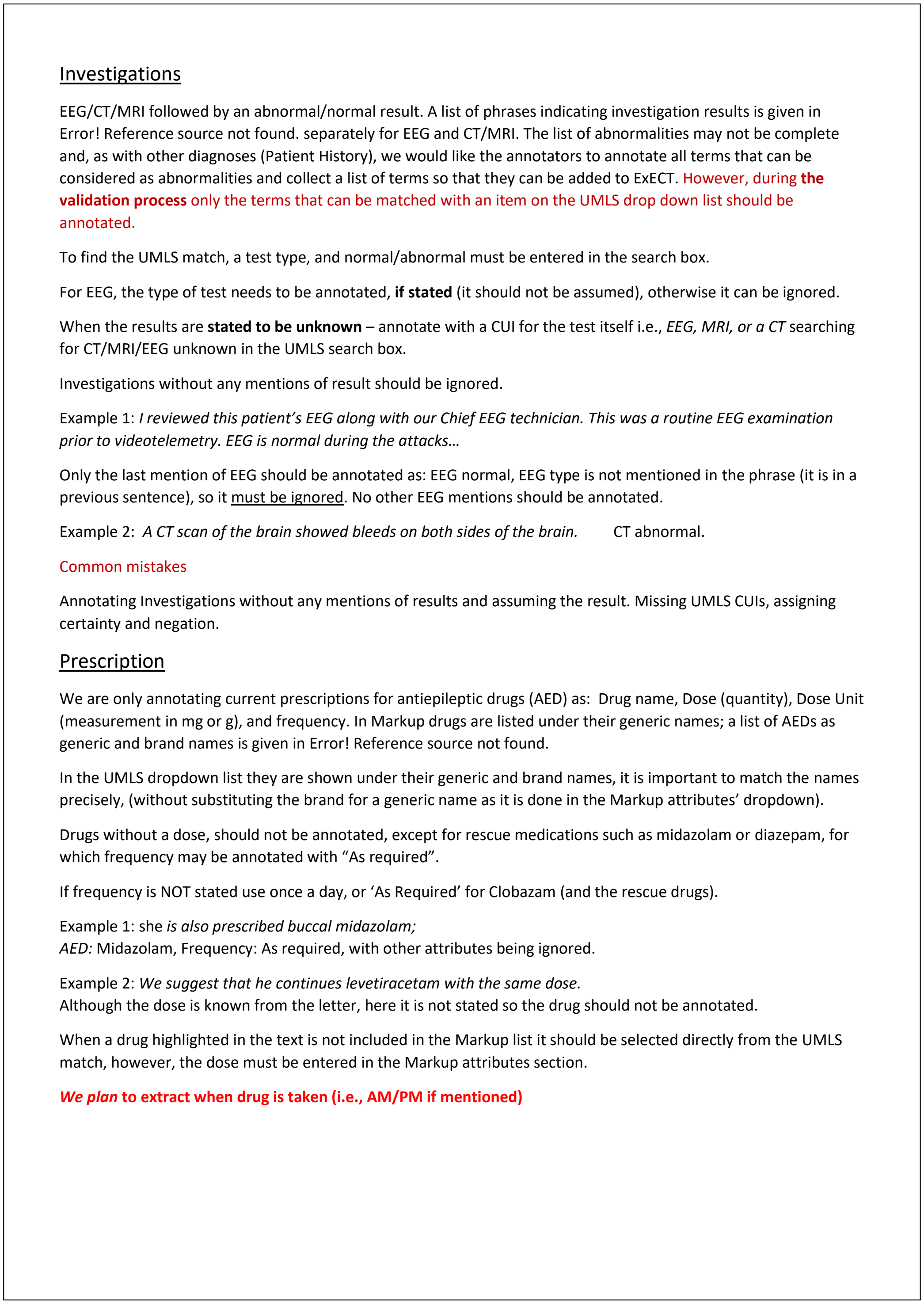
Injuries sustained during or before the time of birth should be annotated here. These have their own CUIs, so it is important to highlight a whole phrase while annotating for the UMLS match to appear.

Normal birth should be annotated, but normal delivery at this stage should not. We are revising the CUI list and may add terms relating to delivery.

Premature Birth – gestational age (grouped) in weeks corresponds to specific UMLS terms / CUIs. A Drop-down in MarkUp would give the appropriate term to search for in the UMLS search box. Levels of prematurity expressed in the number of weeks from full term should be converted to gestational age in weeks.

Example 1: *John was born 8 weeks prematurely;*
BirthHistory: Premature Birth = 32to<37_ModerateToLatePreterm, Certainty = 5, …
The term "moderate to late preterm" should be searched for in the UMLS search box.

Figure B.2: Annotation guidelines page 4



Investigations

EEG/CT/MRI followed by an abnormal/normal result. A list of phrases indicating investigation results is given in Error! Reference source not found. separately for EEG and CT/MRI. The list of abnormalities may not be complete and, as with other diagnoses (Patient History), we would like the annotators to annotate all terms that can be considered as abnormalities and collect a list of terms so that they can be added to ExECT. However, during **the validation process** only the terms that can be matched with an item on the UMLS drop down list should be annotated.

To find the UMLS match, a test type, and normal/abnormal must be entered in the search box.

For EEG, the type of test needs to be annotated, **if stated** (it should not be assumed), otherwise it can be ignored.

When the results are **stated to be unknown** – annotate with a CUI for the test itself i.e., *EEG, MRI, or a CT* searching for CT/MRI/EEG unknown in the UMLS search box.

Investigations without any mentions of result should be ignored.

Example 1: *I reviewed this patient's EEG along with our Chief EEG technician. This was a routine EEG examination prior to videotelemetry. EEG is normal during the attacks…*

Only the last mention of EEG should be annotated as: EEG normal, EEG type is not mentioned in the phrase (it is in a previous sentence), so it must be ignored. No other EEG mentions should be annotated.

Example 2: *A CT scan of the brain showed bleeds on both sides of the brain.* CT abnormal.

Common mistakes

Annotating Investigations without any mentions of results and assuming the result. Missing UMLS CUIs, assigning certainty and negation.

Prescription

We are only annotating current prescriptions for antiepileptic drugs (AED) as: Drug name, Dose (quantity), Dose Unit (measurement in mg or g), and frequency. In Markup drugs are listed under their generic names; a list of AEDs as generic and brand names is given in Error! Reference source not found.

In the UMLS dropdown list they are shown under their generic and brand names, it is important to match the names precisely, (without substituting the brand for a generic name as it is done in the Markup attributes' dropdown).

Drugs without a dose, should not be annotated, except for rescue medications such as midazolam or diazepam, for which frequency may be annotated with "As required".

If frequency is NOT stated use once a day, or 'As Required' for Clobazam (and the rescue drugs).

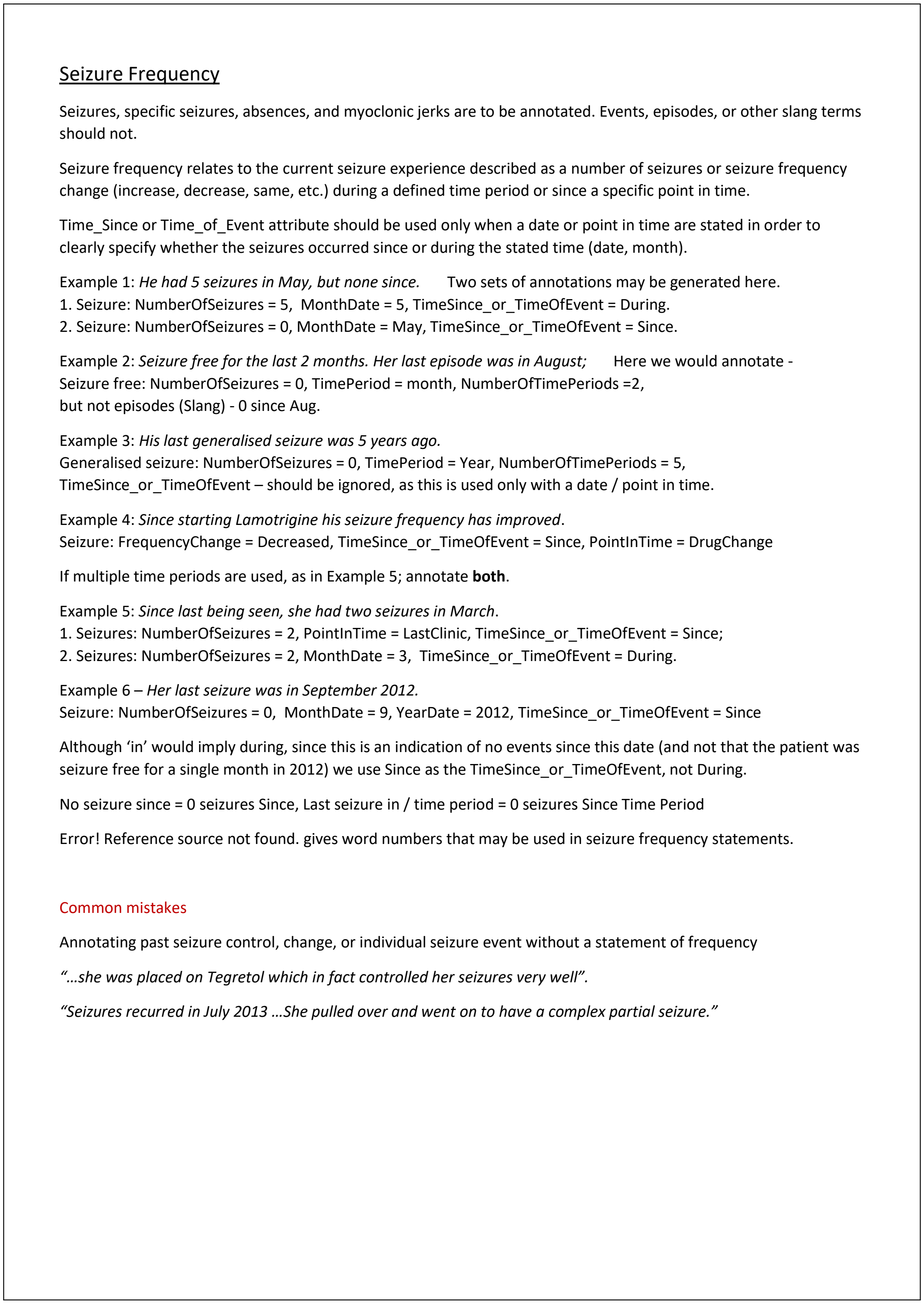
Example 1: she *is also prescribed buccal midazolam;*
AED: Midazolam, Frequency: As required, with other attributes being ignored.

Example 2: *We suggest that he continues levetiracetam with the same dose.*
Although the dose is known from the letter, here it is not stated so the drug should not be annotated.

When a drug highlighted in the text is not included in the Markup list it should be selected directly from the UMLS match, however, the dose must be entered in the Markup attributes section.

***We plan* to extract when drug is taken (i.e., AM/PM if mentioned)**

Figure B.2: Annotation guidelines page 5



Seizure Frequency

Seizures, specific seizures, absences, and myoclonic jerks are to be annotated. Events, episodes, or other slang terms should not.

Seizure frequency relates to the current seizure experience described as a number of seizures or seizure frequency change (increase, decrease, same, etc.) during a defined time period or since a specific point in time.

Time_Since or Time_of_Event attribute should be used only when a date or point in time are stated in order to clearly specify whether the seizures occurred since or during the stated time (date, month).

Example 1: *He had 5 seizures in May, but none since.* Two sets of annotations may be generated here.
1. Seizure: NumberOfSeizures = 5, MonthDate = 5, TimeSince_or_TimeOfEvent = During.
2. Seizure: NumberOfSeizures = 0, MonthDate = May, TimeSince_or_TimeOfEvent = Since.

Example 2: *Seizure free for the last 2 months. Her last episode was in August;* Here we would annotate -
Seizure free: NumberOfSeizures = 0, TimePeriod = month, NumberOfTimePeriods =2,
but not episodes (Slang) - 0 since Aug.

Example 3: *His last generalised seizure was 5 years ago.*
Generalised seizure: NumberOfSeizures = 0, TimePeriod = Year, NumberOfTimePeriods = 5,
TimeSince_or_TimeOfEvent – should be ignored, as this is used only with a date / point in time.

Example 4: *Since starting Lamotrigine his seizure frequency has improved*.
Seizure: FrequencyChange = Decreased, TimeSince_or_TimeOfEvent = Since, PointInTime = DrugChange

If multiple time periods are used, as in Example 5; annotate **both**.

Example 5: *Since last being seen, she had two seizures in March*.
1. Seizures: NumberOfSeizures = 2, PointInTime = LastClinic, TimeSince_or_TimeOfEvent = Since;
2. Seizures: NumberOfSeizures = 2, MonthDate = 3, TimeSince_or_TimeOfEvent = During.

Example 6 – *Her last seizure was in September 2012.*
Seizure: NumberOfSeizures = 0, MonthDate = 9, YearDate = 2012, TimeSince_or_TimeOfEvent = Since

Although 'in' would imply during, since this is an indication of no events since this date (and not that the patient was seizure free for a single month in 2012) we use Since as the TimeSince_or_TimeOfEvent, not During.

No seizure since = 0 seizures Since, Last seizure in / time period = 0 seizures Since Time Period

Error! Reference source not found. gives word numbers that may be used in seizure frequency statements.

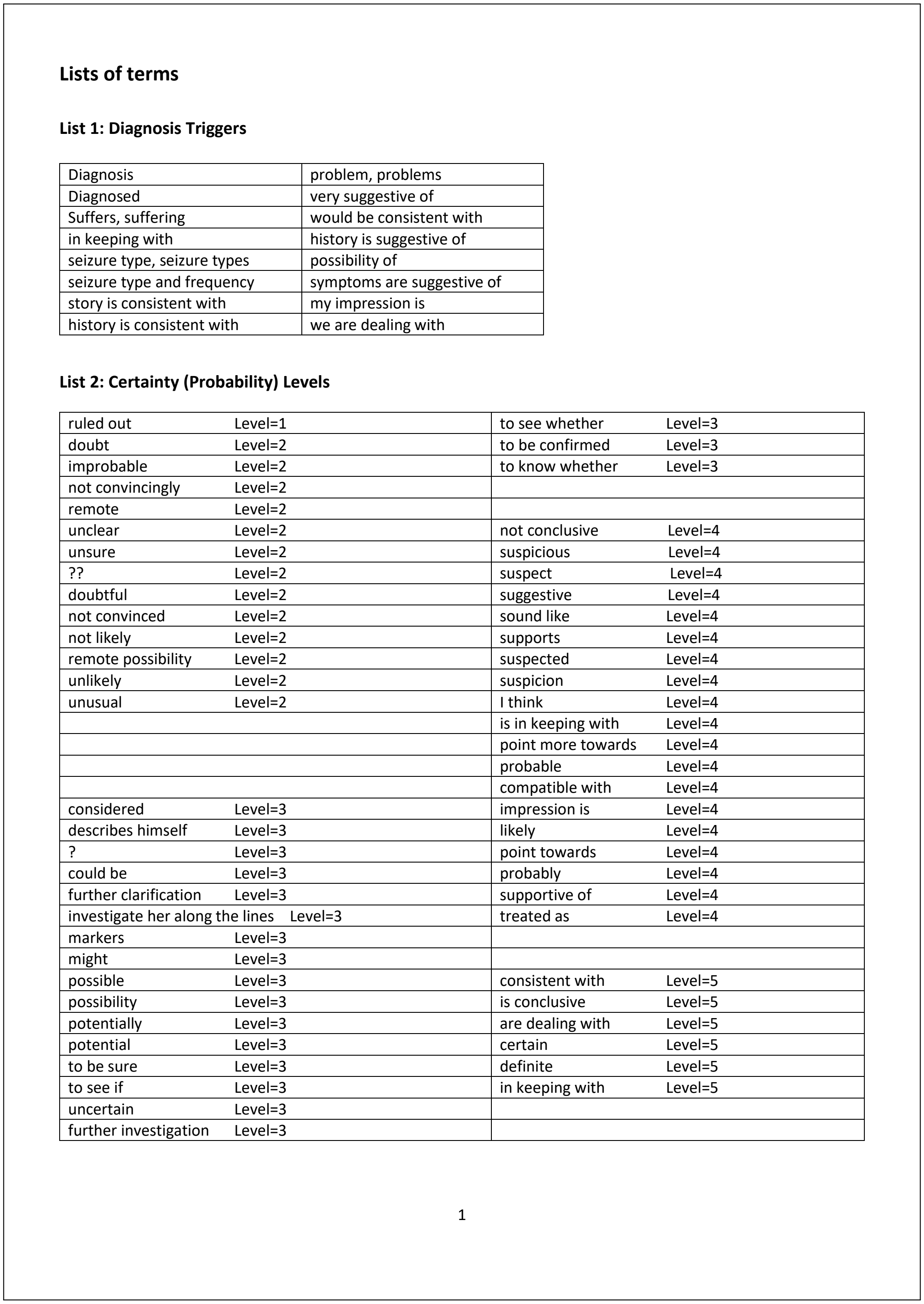
Common mistakes

Annotating past seizure control, change, or individual seizure event without a statement of frequency

"…she was placed on Tegretol which in fact controlled her seizures very well".

"Seizures recurred in July 2013 …She pulled over and went on to have a complex partial seizure."

Figure B.2: Annotation guidelines page 6



Lists of terms

List 1: Diagnosis Triggers

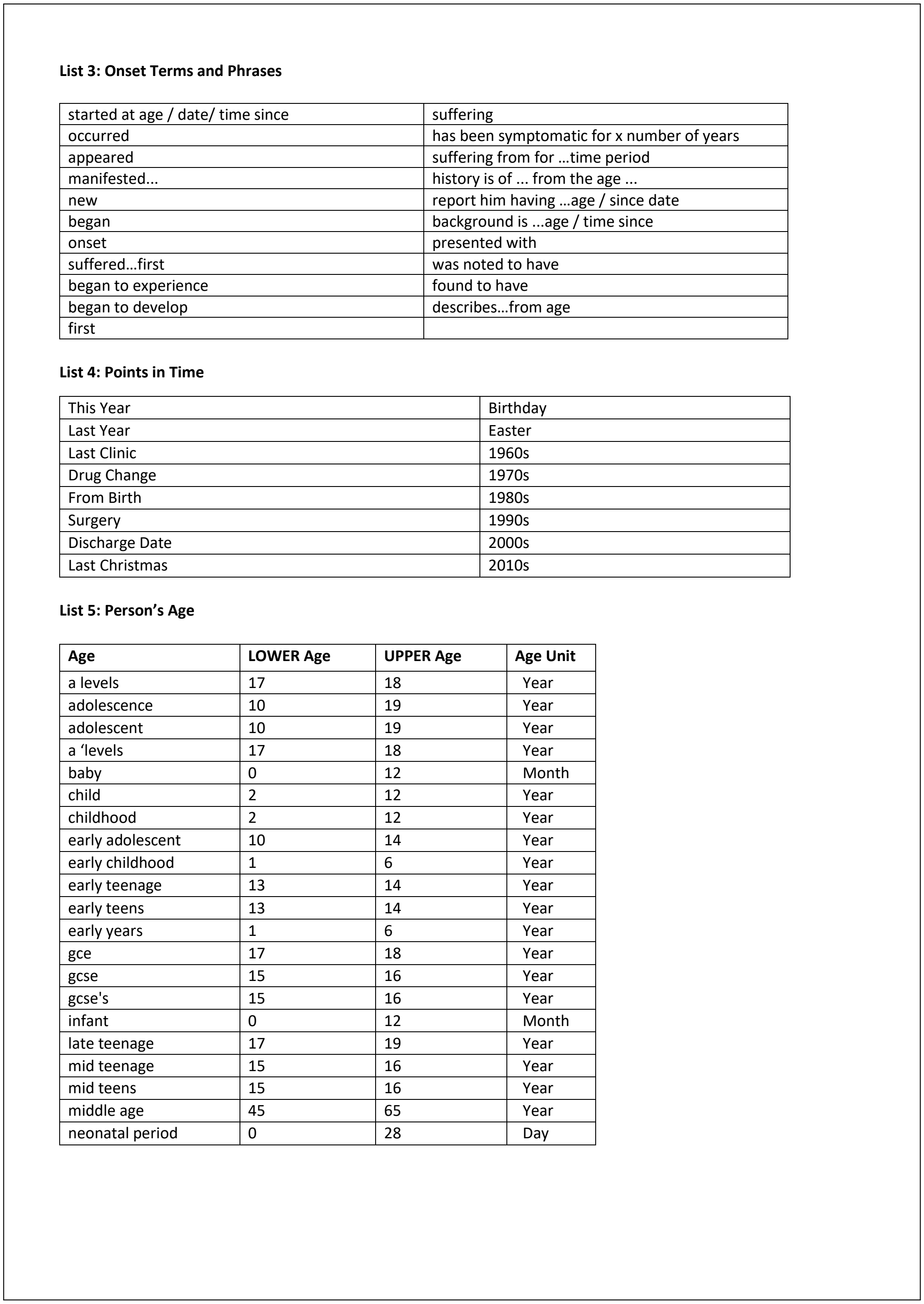
Diagnosis	problem, problems
Diagnosed	very suggestive of
Suffers, suffering	would be consistent with
in keeping with	history is suggestive of
seizure type, seizure types	possibility of
seizure type and frequency	symptoms are suggestive of
story is consistent with	my impression is
history is consistent with	we are dealing with

List 2: Certainty (Probability) Levels

ruled out	Level=1	to see whether	Level=3
doubt	Level=2	to be confirmed	Level=3
improbable	Level=2	to know whether	Level=3
not convincingly	Level=2		
remote	Level=2		
unclear	Level=2	not conclusive	Level=4
unsure	Level=2	suspicious	Level=4
??	Level=2	suspect	Level=4
doubtful	Level=2	suggestive	Level=4
not convinced	Level=2	sound like	Level=4
not likely	Level=2	supports	Level=4
remote possibility	Level=2	suspected	Level=4
unlikely	Level=2	suspicion	Level=4
unusual	Level=2	I think	Level=4
		is in keeping with	Level=4
		point more towards	Level=4
		probable	Level=4
		compatible with	Level=4
considered	Level=3	impression is	Level=4
describes himself	Level=3	likely	Level=4
?	Level=3	point towards	Level=4
could be	Level=3	probably	Level=4
further clarification	Level=3	supportive of	Level=4
investigate her along the lines	Level=3	treated as	Level=4
markers	Level=3		
might	Level=3		
possible	Level=3	consistent with	Level=5
possibility	Level=3	is conclusive	Level=5
potentially	Level=3	are dealing with	Level=5
potential	Level=3	certain	Level=5
to be sure	Level=3	definite	Level=5
to see if	Level=3	in keeping with	Level=5
uncertain	Level=3		
further investigation	Level=3		

1

Figure B.2: Annotation guidelines page 7



List 3: Onset Terms and Phrases

started at age / date/ time since	suffering
occurred	has been symptomatic for x number of years
appeared	suffering from for …time period
manifested...	history is of … from the age …
new	report him having …age / since date
began	background is …age / time since
onset	presented with
suffered…first	was noted to have
began to experience	found to have
began to develop	describes…from age
first	

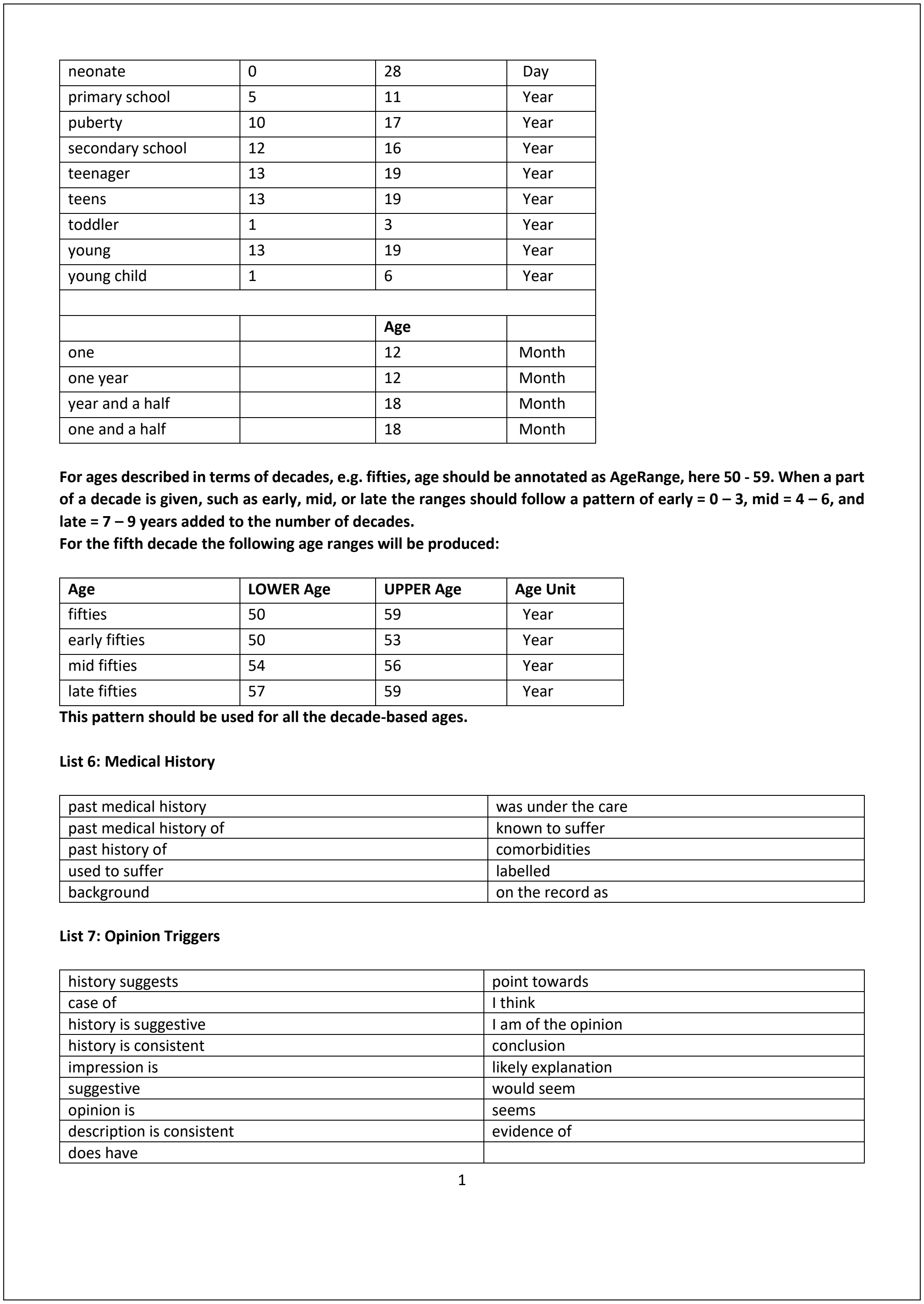
List 4: Points in Time

This Year	Birthday
Last Year	Easter
Last Clinic	1960s
Drug Change	1970s
From Birth	1980s
Surgery	1990s
Discharge Date	2000s
Last Christmas	2010s

List 5: Person’s Age

Age	LOWER Age	UPPER Age	Age Unit
a levels	17	18	Year
adolescence	10	19	Year
adolescent	10	19	Year
a ‘levels	17	18	Year
baby	0	12	Month
child	2	12	Year
childhood	2	12	Year
early adolescent	10	14	Year
early childhood	1	6	Year
early teenage	13	14	Year
early teens	13	14	Year
early years	1	6	Year
gce	17	18	Year
gcse	15	16	Year
gcse's	15	16	Year
infant	0	12	Month
late teenage	17	19	Year
mid teenage	15	16	Year
mid teens	15	16	Year
middle age	45	65	Year
neonatal period	0	28	Day

Figure B.2: Annotation guidelines page 8



neonate	0	28	Day
primary school	5	11	Year
puberty	10	17	Year
secondary school	12	16	Year
teenager	13	19	Year
teens	13	19	Year
toddler	1	3	Year
young	13	19	Year
young child	1	6	Year
		Age	
one		12	Month
one year		12	Month
year and a half		18	Month
one and a half		18	Month

For ages described in terms of decades, e.g. fifties, age should be annotated as AgeRange, here 50 - 59. When a part of a decade is given, such as early, mid, or late the ranges should follow a pattern of early = 0 – 3, mid = 4 – 6, and late = 7 – 9 years added to the number of decades.
For the fifth decade the following age ranges will be produced:

Age	**LOWER Age**	**UPPER Age**	**Age Unit**
fifties	50	59	Year
early fifties	50	53	Year
mid fifties	54	56	Year
late fifties	57	59	Year

This pattern should be used for all the decade-based ages.

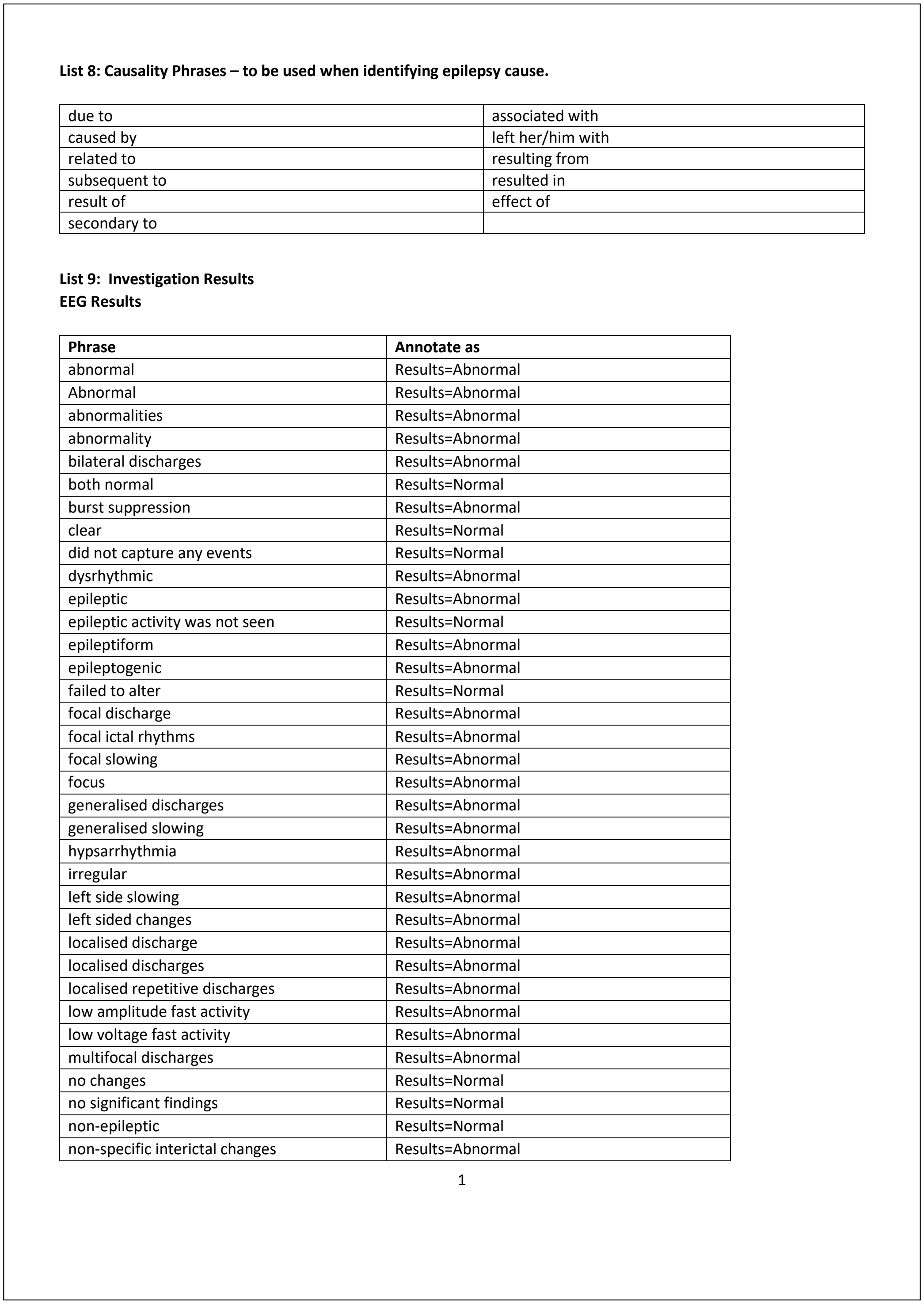
List 6: Medical History

past medical history	was under the care
past medical history of	known to suffer
past history of	comorbidities
used to suffer	labelled
background	on the record as

List 7: Opinion Triggers

history suggests	point towards
case of	I think
history is suggestive	I am of the opinion
history is consistent	conclusion
impression is	likely explanation
suggestive	would seem
opinion is	seems
description is consistent	evidence of
does have	

Figure B.2: Annotation guidelines page 9



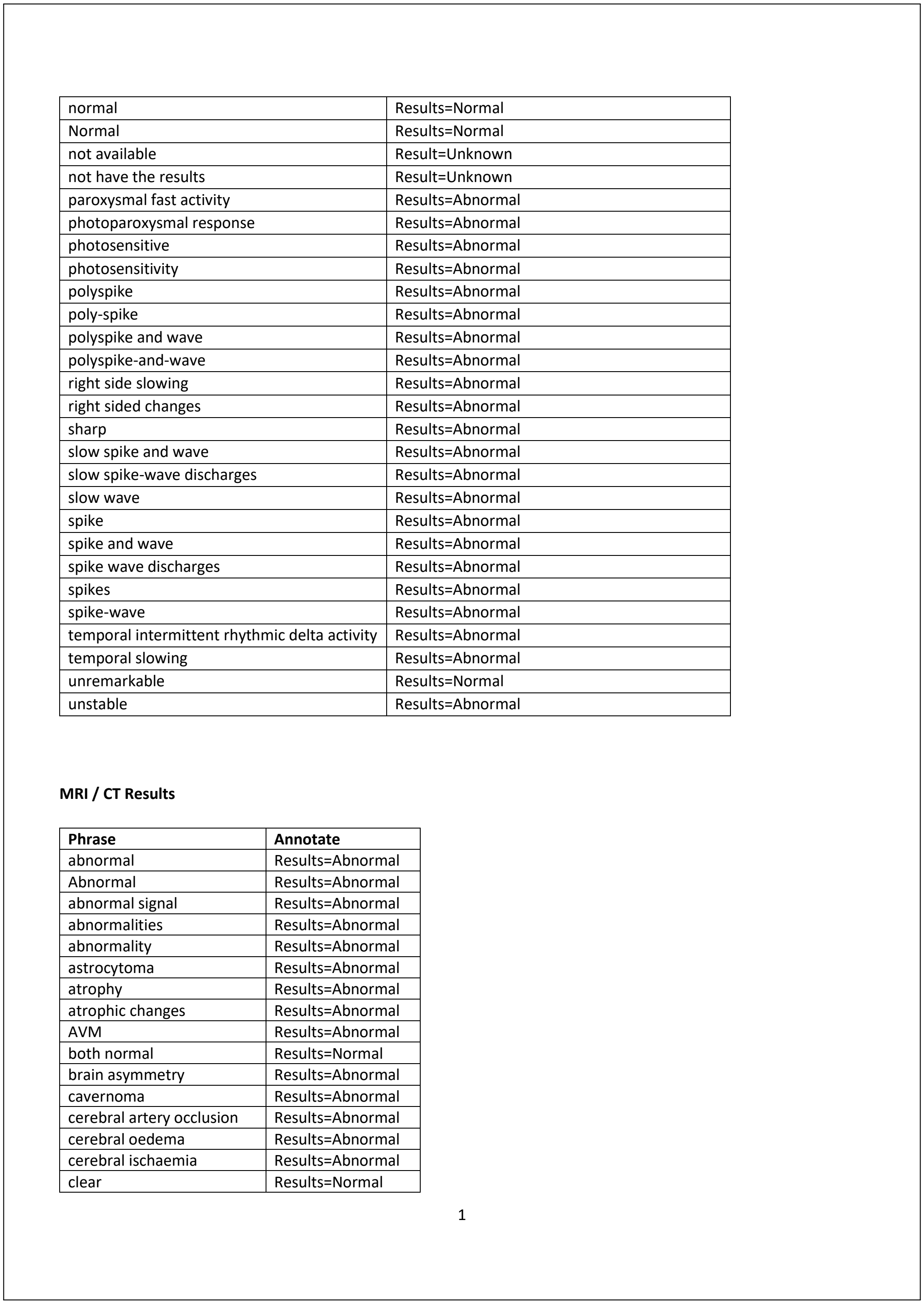
List 8: Causality Phrases – to be used when identifying epilepsy cause.

due to	associated with
caused by	left her/him with
related to	resulting from
subsequent to	resulted in
result of	effect of
secondary to	

List 9: Investigation Results
EEG Results

Phrase	Annotate as
abnormal	Results=Abnormal
Abnormal	Results=Abnormal
abnormalities	Results=Abnormal
abnormality	Results=Abnormal
bilateral discharges	Results=Abnormal
both normal	Results=Normal
burst suppression	Results=Abnormal
clear	Results=Normal
did not capture any events	Results=Normal
dysrhythmic	Results=Abnormal
epileptic	Results=Abnormal
epileptic activity was not seen	Results=Normal
epileptiform	Results=Abnormal
epileptogenic	Results=Abnormal
failed to alter	Results=Normal
focal discharge	Results=Abnormal
focal ictal rhythms	Results=Abnormal
focal slowing	Results=Abnormal
focus	Results=Abnormal
generalised discharges	Results=Abnormal
generalised slowing	Results=Abnormal
hypsarrhythmia	Results=Abnormal
irregular	Results=Abnormal
left side slowing	Results=Abnormal
left sided changes	Results=Abnormal
localised discharge	Results=Abnormal
localised discharges	Results=Abnormal
localised repetitive discharges	Results=Abnormal
low amplitude fast activity	Results=Abnormal
low voltage fast activity	Results=Abnormal
multifocal discharges	Results=Abnormal
no changes	Results=Normal
no significant findings	Results=Normal
non-epileptic	Results=Normal
non-specific interictal changes	Results=Abnormal

Figure B.2: Annotation guidelines page 10



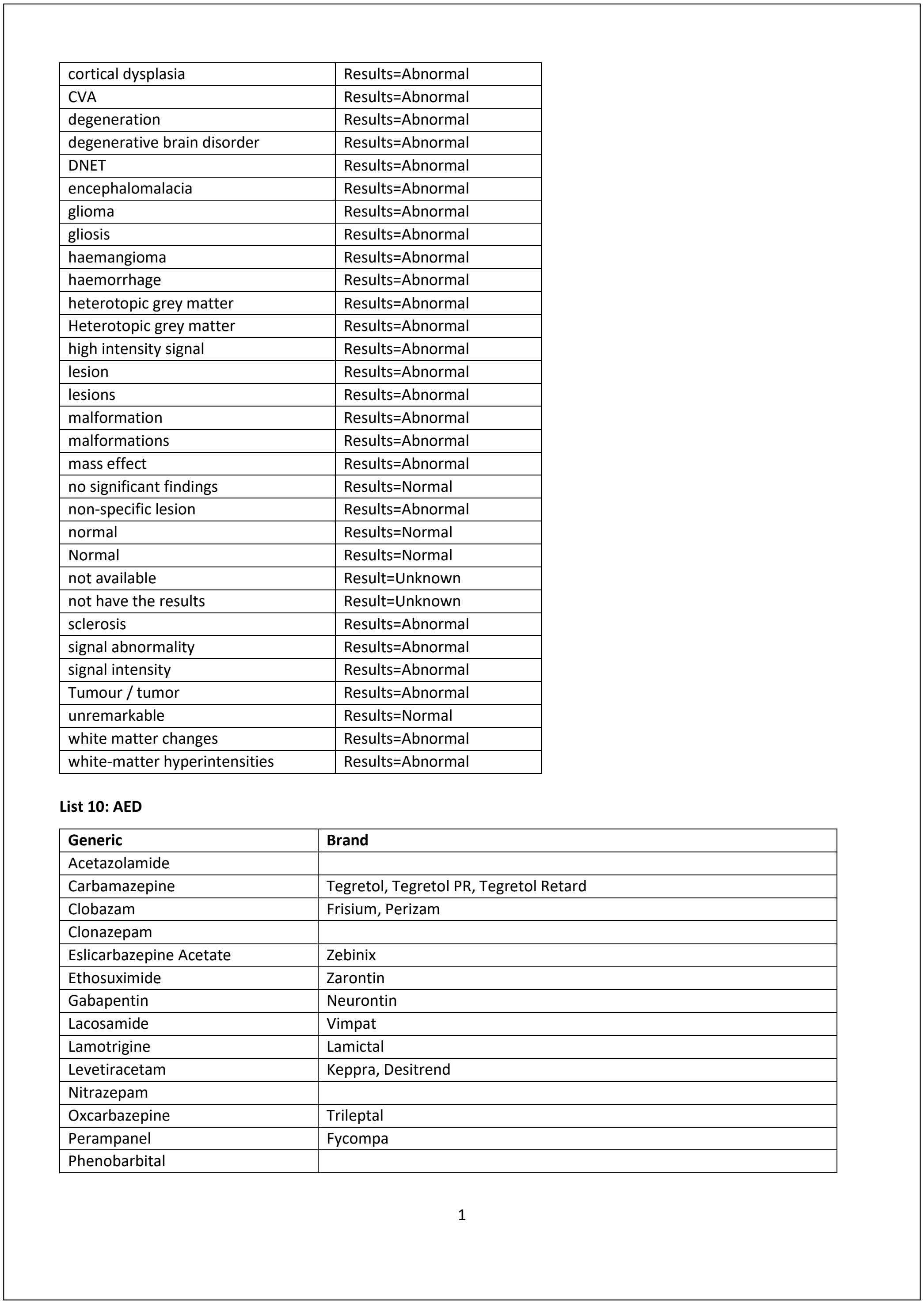
normal	Results=Normal
Normal	Results=Normal
not available	Result=Unknown
not have the results	Result=Unknown
paroxysmal fast activity	Results=Abnormal
photoparoxysmal response	Results=Abnormal
photosensitive	Results=Abnormal
photosensitivity	Results=Abnormal
polyspike	Results=Abnormal
poly-spike	Results=Abnormal
polyspike and wave	Results=Abnormal
polyspike-and-wave	Results=Abnormal
right side slowing	Results=Abnormal
right sided changes	Results=Abnormal
sharp	Results=Abnormal
slow spike and wave	Results=Abnormal
slow spike-wave discharges	Results=Abnormal
slow wave	Results=Abnormal
spike	Results=Abnormal
spike and wave	Results=Abnormal
spike wave discharges	Results=Abnormal
spikes	Results=Abnormal
spike-wave	Results=Abnormal
temporal intermittent rhythmic delta activity	Results=Abnormal
temporal slowing	Results=Abnormal
unremarkable	Results=Normal
unstable	Results=Abnormal

MRI / CT Results

Phrase	Annotate
abnormal	Results=Abnormal
Abnormal	Results=Abnormal
abnormal signal	Results=Abnormal
abnormalities	Results=Abnormal
abnormality	Results=Abnormal
astrocytoma	Results=Abnormal
atrophy	Results=Abnormal
atrophic changes	Results=Abnormal
AVM	Results=Abnormal
both normal	Results=Normal
brain asymmetry	Results=Abnormal
cavernoma	Results=Abnormal
cerebral artery occlusion	Results=Abnormal
cerebral oedema	Results=Abnormal
cerebral ischaemia	Results=Abnormal
clear	Results=Normal

1

Figure B.2: Annotation guidelines page 11



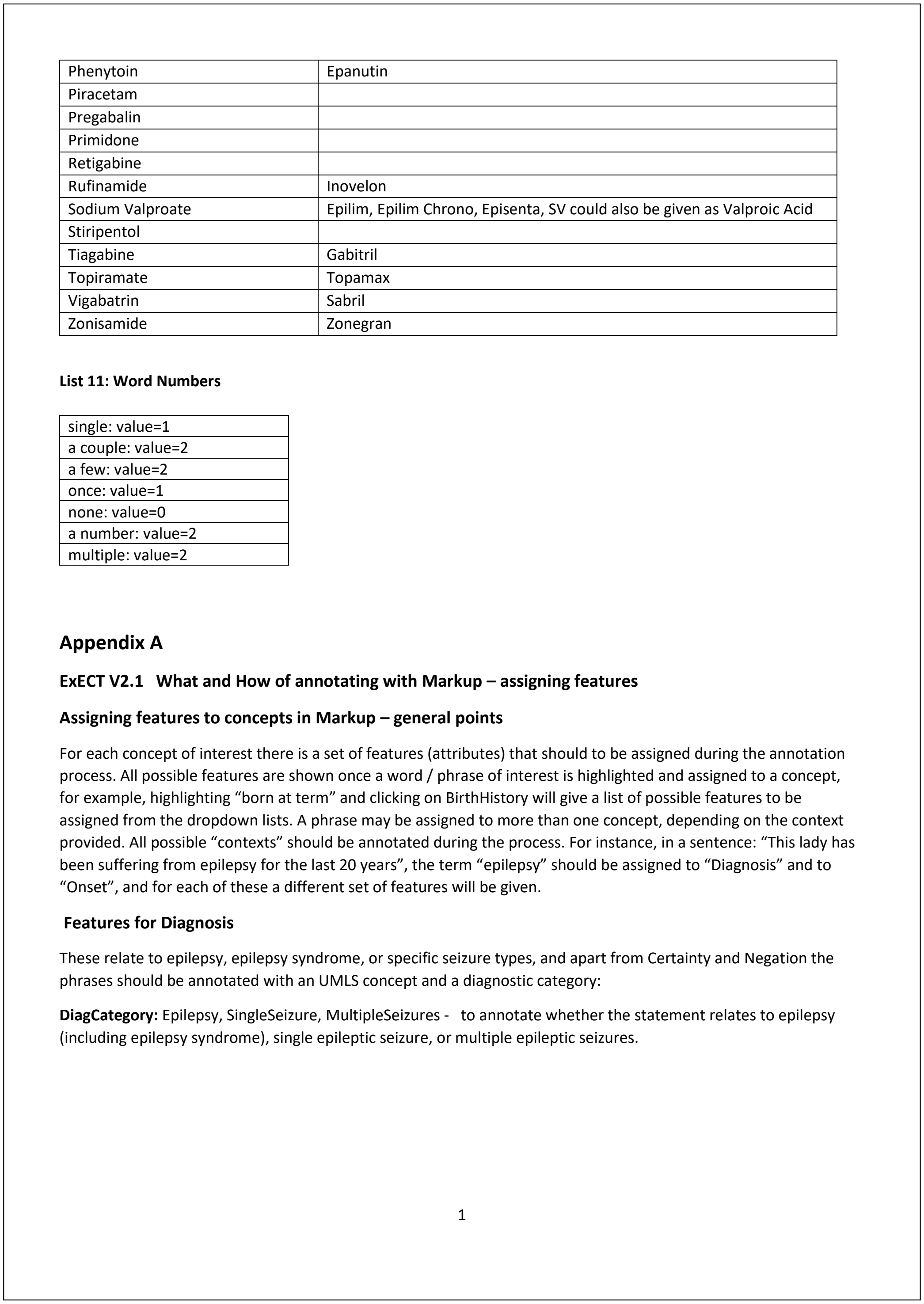
cortical dysplasia	Results=Abnormal
CVA	Results=Abnormal
degeneration	Results=Abnormal
degenerative brain disorder	Results=Abnormal
DNET	Results=Abnormal
encephalomalacia	Results=Abnormal
glioma	Results=Abnormal
gliosis	Results=Abnormal
haemangioma	Results=Abnormal
haemorrhage	Results=Abnormal
heterotopic grey matter	Results=Abnormal
Heterotopic grey matter	Results=Abnormal
high intensity signal	Results=Abnormal
lesion	Results=Abnormal
lesions	Results=Abnormal
malformation	Results=Abnormal
malformations	Results=Abnormal
mass effect	Results=Abnormal
no significant findings	Results=Normal
non-specific lesion	Results=Abnormal
normal	Results=Normal
Normal	Results=Normal
not available	Result=Unknown
not have the results	Result=Unknown
sclerosis	Results=Abnormal
signal abnormality	Results=Abnormal
signal intensity	Results=Abnormal
Tumour / tumor	Results=Abnormal
unremarkable	Results=Normal
white matter changes	Results=Abnormal
white-matter hyperintensities	Results=Abnormal

List 10: AED

Generic	**Brand**
Acetazolamide	
Carbamazepine	Tegretol, Tegretol PR, Tegretol Retard
Clobazam	Frisium, Perizam
Clonazepam	
Eslicarbazepine Acetate	Zebinix
Ethosuximide	Zarontin
Gabapentin	Neurontin
Lacosamide	Vimpat
Lamotrigine	Lamictal
Levetiracetam	Keppra, Desitrend
Nitrazepam	
Oxcarbazepine	Trileptal
Perampanel	Fycompa
Phenobarbital	

1

Figure B.2: Annotation guidelines page 12



Phenytoin	Epanutin
Piracetam	
Pregabalin	
Primidone	
Retigabine	
Rufinamide	Inovelon
Sodium Valproate	Epilim, Epilim Chrono, Episenta, SV could also be given as Valproic Acid
Stiripentol	
Tiagabine	Gabitril
Topiramate	Topamax
Vigabatrin	Sabril
Zonisamide	Zonegran

List 11: Word Numbers

single: value=1
a couple: value=2
a few: value=2
once: value=1
none: value=0
a number: value=2
multiple: value=2

Appendix A

ExECT V2.1 What and How of annotating with Markup – assigning features

Assigning features to concepts in Markup – general points

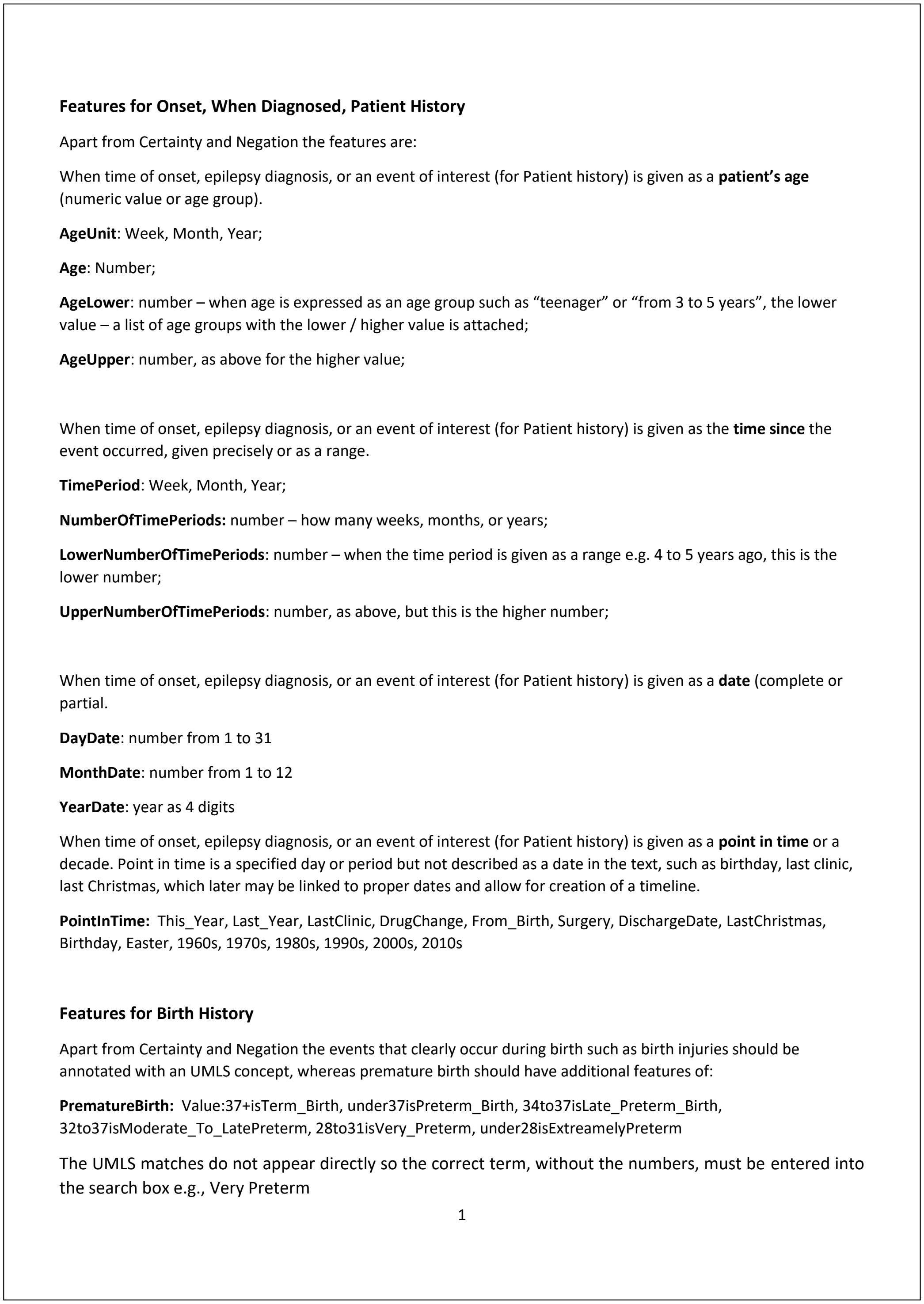
For each concept of interest there is a set of features (attributes) that should to be assigned during the annotation process. All possible features are shown once a word / phrase of interest is highlighted and assigned to a concept, for example, highlighting "born at term" and clicking on BirthHistory will give a list of possible features to be assigned from the dropdown lists. A phrase may be assigned to more than one concept, depending on the context provided. All possible "contexts" should be annotated during the process. For instance, in a sentence: "This lady has been suffering from epilepsy for the last 20 years", the term "epilepsy" should be assigned to "Diagnosis" and to "Onset", and for each of these a different set of features will be given.

Features for Diagnosis

These relate to epilepsy, epilepsy syndrome, or specific seizure types, and apart from Certainty and Negation the phrases should be annotated with an UMLS concept and a diagnostic category:

DiagCategory: Epilepsy, SingleSeizure, MultipleSeizures - to annotate whether the statement relates to epilepsy (including epilepsy syndrome), single epileptic seizure, or multiple epileptic seizures.

Figure B.2: Annotation guidelines page 13



Features for Onset, When Diagnosed, Patient History

Apart from Certainty and Negation the features are:

When time of onset, epilepsy diagnosis, or an event of interest (for Patient history) is given as a **patient's age** (numeric value or age group).

AgeUnit: Week, Month, Year;

Age: Number;

AgeLower: number – when age is expressed as an age group such as "teenager" or "from 3 to 5 years", the lower value – a list of age groups with the lower / higher value is attached;

AgeUpper: number, as above for the higher value;

When time of onset, epilepsy diagnosis, or an event of interest (for Patient history) is given as the **time since** the event occurred, given precisely or as a range.

TimePeriod: Week, Month, Year;

NumberOfTimePeriods: number – how many weeks, months, or years;

LowerNumberOfTimePeriods: number – when the time period is given as a range e.g. 4 to 5 years ago, this is the lower number;

UpperNumberOfTimePeriods: number, as above, but this is the higher number;

When time of onset, epilepsy diagnosis, or an event of interest (for Patient history) is given as a **date** (complete or partial.

DayDate: number from 1 to 31

MonthDate: number from 1 to 12

YearDate: year as 4 digits

When time of onset, epilepsy diagnosis, or an event of interest (for Patient history) is given as a **point in time** or a decade. Point in time is a specified day or period but not described as a date in the text, such as birthday, last clinic, last Christmas, which later may be linked to proper dates and allow for creation of a timeline.

PointInTime: This_Year, Last_Year, LastClinic, DrugChange, From_Birth, Surgery, DischargeDate, LastChristmas, Birthday, Easter, 1960s, 1970s, 1980s, 1990s, 2000s, 2010s

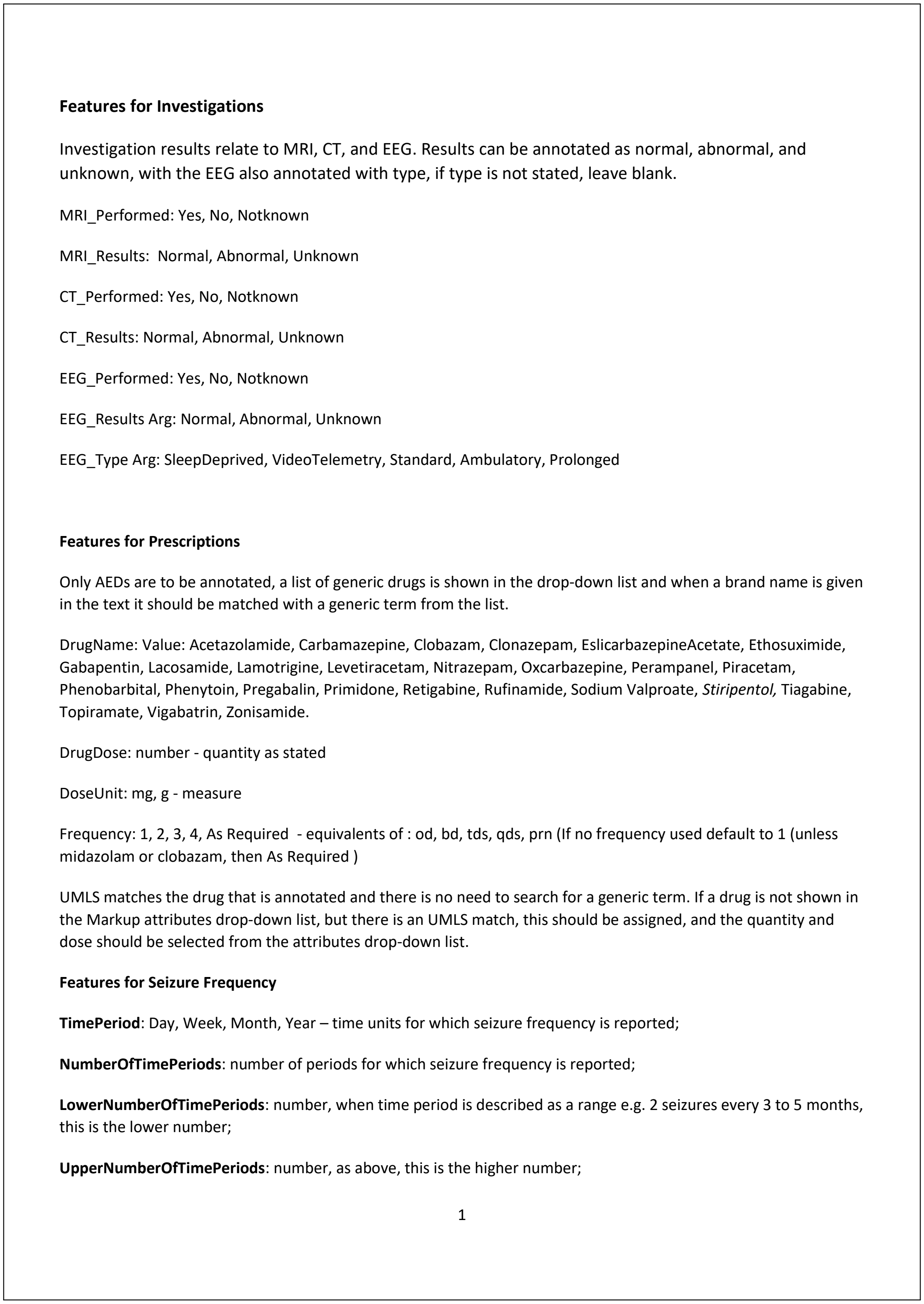
Features for Birth History

Apart from Certainty and Negation the events that clearly occur during birth such as birth injuries should be annotated with an UMLS concept, whereas premature birth should have additional features of:

PrematureBirth: Value:37+isTerm_Birth, under37isPreterm_Birth, 34to37isLate_Preterm_Birth, 32to37isModerate_To_LatePreterm, 28to31isVery_Preterm, under28isExtreamelyPreterm

The UMLS matches do not appear directly so the correct term, without the numbers, must be entered into the search box e.g., Very Preterm

Figure B.2: Annotation guidelines page 14



Features for Investigations

Investigation results relate to MRI, CT, and EEG. Results can be annotated as normal, abnormal, and unknown, with the EEG also annotated with type, if type is not stated, leave blank.

MRI_Performed: Yes, No, Notknown

MRI_Results: Normal, Abnormal, Unknown

CT_Performed: Yes, No, Notknown

CT_Results: Normal, Abnormal, Unknown

EEG_Performed: Yes, No, Notknown

EEG_Results Arg: Normal, Abnormal, Unknown

EEG_Type Arg: SleepDeprived, VideoTelemetry, Standard, Ambulatory, Prolonged

Features for Prescriptions

Only AEDs are to be annotated, a list of generic drugs is shown in the drop-down list and when a brand name is given in the text it should be matched with a generic term from the list.

DrugName: Value: Acetazolamide, Carbamazepine, Clobazam, Clonazepam, EslicarbazepineAcetate, Ethosuximide, Gabapentin, Lacosamide, Lamotrigine, Levetiracetam, Nitrazepam, Oxcarbazepine, Perampanel, Piracetam, Phenobarbital, Phenytoin, Pregabalin, Primidone, Retigabine, Rufinamide, Sodium Valproate, *Stiripentol,* Tiagabine, Topiramate, Vigabatrin, Zonisamide.

DrugDose: number - quantity as stated

DoseUnit: mg, g - measure

Frequency: 1, 2, 3, 4, As Required - equivalents of : od, bd, tds, qds, prn (If no frequency used default to 1 (unless midazolam or clobazam, then As Required)

UMLS matches the drug that is annotated and there is no need to search for a generic term. If a drug is not shown in the Markup attributes drop-down list, but there is an UMLS match, this should be assigned, and the quantity and dose should be selected from the attributes drop-down list.

Features for Seizure Frequency

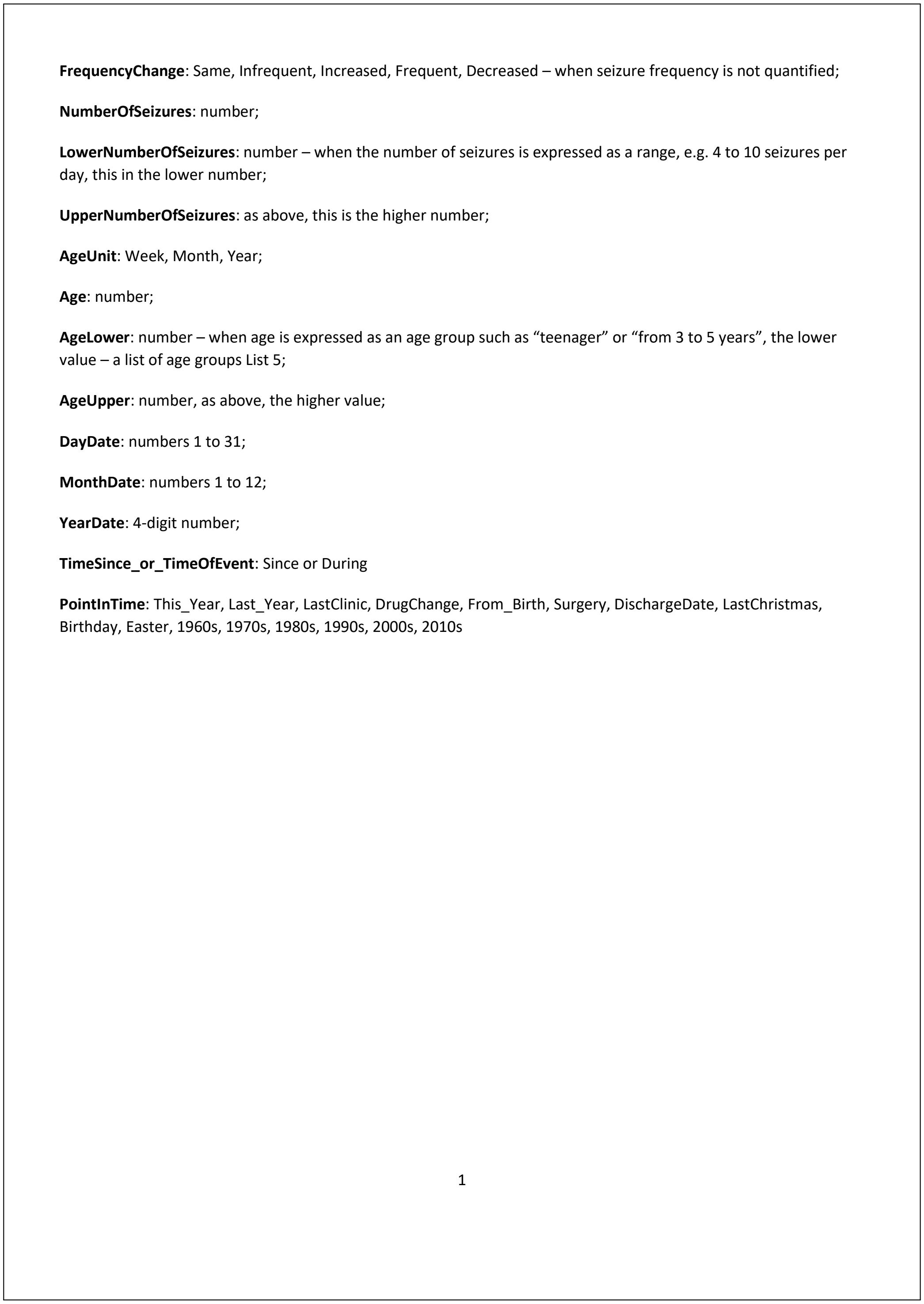
TimePeriod: Day, Week, Month, Year – time units for which seizure frequency is reported;

NumberOfTimePeriods: number of periods for which seizure frequency is reported;

LowerNumberOfTimePeriods: number, when time period is described as a range e.g. 2 seizures every 3 to 5 months, this is the lower number;

UpperNumberOfTimePeriods: number, as above, this is the higher number;

Figure B.2: Annotation guidelines page 15



FrequencyChange: Same, Infrequent, Increased, Frequent, Decreased – when seizure frequency is not quantified;

NumberOfSeizures: number;

LowerNumberOfSeizures: number – when the number of seizures is expressed as a range, e.g. 4 to 10 seizures per day, this in the lower number;

UpperNumberOfSeizures: as above, this is the higher number;

AgeUnit: Week, Month, Year;

Age: number;

AgeLower: number – when age is expressed as an age group such as “teenager” or “from 3 to 5 years”, the lower value – a list of age groups List 5;

AgeUpper: number, as above, the higher value;

DayDate: numbers 1 to 31;

MonthDate: numbers 1 to 12;

YearDate: 4-digit number;

TimeSince_or_TimeOfEvent: Since or During

PointInTime: This_Year, Last_Year, LastClinic, DrugChange, From_Birth, Surgery, DischargeDate, LastChristmas, Birthday, Easter, 1960s, 1970s, 1980s, 1990s, 2000s, 2010s

Figure B.2: Annotation guidelines page 16

Appendix C

Epi25 Cohort processing – output validation and analysis

C.1 Diagnosis in the SNB database

C.2 Epi25 Onset processing

C.3 Seizure frequency processing

C.4 Prescription processing (fragment)

```r

# extracting diagnosis from the biobank based on epilepsy diagnosis and on seizure type against a list of CUIs used in ExECT
# selects SNB individuals with focal epilepsy based on epilepsy and seizure type
SNB_diag_focal <- SNB_Diag_Seizures %>%
inner_join(FocalCUIList) %>%
distinct(SYSTEM_ID,EpilepsyType) # groups by SYSTEM_ID

# selects SNB individuals with generalised epilepsy based on epilepsy and seizure type
SNB_diag_generalised <- SNB_Diag_Seizures %>%
inner_join(GeneralisedCUIList) %>%
distinct(SYSTEM_ID, EpilepsyType) # groups by SYSTEM_ID

# to check that nobody had two types of diagnosis we combine the tables and select distinct System_ID
SNB_diag_focal_generalised <- full_join(SNB_diag_focal, SNB_diag_generalised)

Distinct <- SNB_diag_focal_generalised %>%
    distinct(SYSTEM_ID)
# renamig EpilepsyType to SNB_EpilepsyType for validation of the Epi25 set
SNB_diag_focal_generalised <- rename(SNB_diag_focal_generalised, SNB_EpilepsyType = EpilepsyType )


```

Figure C.1: r script creating single epilepsy diagnosis table from from the SNB database record of epilepsy and seizure using ExECT list od epilepsy and seizure terms

```r
#reading in Onset Output that has been linked to record date
Epi25_Onset_DOC <- read_delim("C:/Users/Beata/Documents/Epi25letters/ExECTOutput/Epi25_Onset_DOC.csv", delim = ":", escape_double = FALSE, trim_ws = TRUE)
# identifying individuals
Epi25OnsetPeople <- Epi25_Onset_DOC %>%
    distinct(SYSTEM_ID)

# bringing in date of birth extracted by IDEx----
DoB <-read_csv("C:/Users/Beata/Documents/Epi25letters/IDExOutput/DoB.csv", delim = ":", escape_double = FALSE, trim_ws = TRUE)
# making sure date of birth in in date format
DoB$DATE_OF_BIRTH <- as.Date(DoB$DATE_OF_BIRTH, "%d/%m/%Y")
# extracting date of birth
DoBonly <- DoB %>%
  select(DATE_OF_BIRTH, SYSTEM_ID) %>%
  group_by(SYSTEM_ID) %>%
  unique()

#Adding date of birth to the output for onset
Epi25Onset <- left_join(Epi25_Onset_DOC, DoBonly)

#converting character values to numbers and dates where needed
Epi25Onset$NoTP <- as.numeric(Epi25Onset$NoTP)
Epi25Onset$Age <- as.numeric(Epi25Onset$Age) # age
Epi25Onset$DATEREC <- as.Date(Epi25Onset$DATEREC, "%d/%m/%Y")  # Record date
Epi25Onset$DATE_OF_BIRTH <- as.Date(Epi25Onset$DATE_OF_BIRTH, "%d/%m/%Y") # Date of birth
Epi25Onset$YD <- as.numeric(Epi25Onset$YD) # convert Year date to numeric
Epi25Onset$MD <- as.numeric(Epi25Onset$MD) # convert Month date to numeric
# OnsetDate based on separate DD,MD, and YD columns ----
# converting numerical values given as days (DD) and months (MD) to two figure format
Epi25Onset <- Epi25Onset %>% mutate(DD = ifelse(!is.na(DD) & as.numeric(DD) < 10,paste("0", DD, sep = "") ,DD)) # convert from 9 to 09 for date
Epi25Onset<- Epi25Onset %>% mutate(MD = ifelse(!is.na(MD) & as.numeric(MD) < 10,paste("0", MD, sep = "") ,MD))  # time periods to days
Epi25Onset <- Epi25Onset %>% mutate(Days = case_when(TP == "Day" ~ 1,
TP == "Week" ~ 7,
TP == "Month" ~ 30,
TP == "Year" ~ 365))
Epi25Onset <- Epi25Onset %>% mutate(TPinDays = case_when(TP == 'Year' ~ NoTP*Days))
# when diffrent date elements are given in the output
Epi25Onset <- Epi25Onset %>% mutate(OnsetDate = ifelse(!is.na(YD) & !is.na(MD) & !is.na(DD) , paste(DD, MD, YD, sep = "/"),
ifelse(!is.na(YD) & !is.na(MD) & is.na(DD), paste("01", MD, YD, sep = "/"),
ifelse(!is.na(YD) & is.na(MD) & is.na(DD), paste("01","01", YD, sep = "/"),
ifelse(is.na(YD) & !is.na(MD) & is.na(DD), paste("01", MD, ifelse(as.numeric(MD) < as.numeric(format(DATEREC, "%m")), format(DATEREC, "%Y"), as.numeric(format(DATEREC, "%Y"))-30), sep = "/"),
ifelse(is.na(YD) & !is.na(MD) & !is.na(DD), paste(DD, MD, ifelse(as.numeric(MD) <  as.numeric(format(DATEREC, "%m")), format(DATEREC, "%Y"), as.numeric(format(DATEREC, "%Y")) -1), sep = "/"), NA))))), .after = "DATEREC")#As above but with day also (in above set the 1st of month)

```

Figure C.2: R script converting `time since', ‘date when', and ‘age'into a single common measure of age of onset.

```r
#when onset was x time periods from clinic date, calculating date of onset2

Epi25Onset <- Epi25Onset %>% mutate(OnsetDate2  = case_when(!is.na(TP) & !is.na(TPinDays) ~ DATEREC - TPinDays),.after = "OnsetDate") #creating another onset date when onset is expressed as time since onset

Epi25Onset <- Epi25Onset %>% mutate(OnsetDate3 = case_when(is.na(OnsetDate) ~ OnsetDate2,
!is.na(OnsetDate) & is.na(OnsetDate2) ~ OnsetDate),.after = "OnsetDate2") # added values from OnsetDate2 to OnsetDate

# When age is given as a range take the lower value as AGEL

Epi25Onset <- Epi25Onset %>% mutate(Age1 = case_when(is.na(Age) & !is.na(AgeL) ~ AgeL), as.numeric(Age1))

Epi25Onset$Age1 <- as.numeric(Epi25Onset$Age1)

#calculated age of onset in days
Epi25Onset <- Epi25Onset %>% mutate(Age3 = case_when(!is.na(OnsetDate3) ~ OnsetDate3-DATE_OF_BIRTH), as.numeric(Age3))
# calculating age in years from difference in days between date of birth and onset date
Epi25Onset <- Epi25Onset %>% mutate(Age4 = as.numeric(Age3)/365)
Epi25Onset$Age4 <- round(Epi25Onset$Age4)

#Final  ONSET selection----
Epi25Onset <- Epi25Onset %>%
  mutate(OnsetAge = case_when(!is.na(Age) ~ Age,
  is.na(Age) & !is.na(Age1) ~ Age1,
  is.na(Age1) & !is.na(Age4) ~ Age4))

#Creating separate columns for onset in years and onset in months to match  SNB format
Epi25Onset <- Epi25Onset %>%
  mutate(AGE_ONSET_Y = case_when(AgeUnit == "Year" ~ OnsetAge, as.numeric(Age3) > 365 ~ OnsetAge))
Epi25Onset <- Epi25Onset %>%
  mutate(AGE_ONSET_M = case_when(AgeUnit == "Month" ~ OnsetAge))

Epi25OnsetFinal <- Epi25Onset %>%
  select(SYSTEM_ID, DATEREC, CUI, PREF,AGE_ONSET_Y, AGE_ONSET_M ) %>%
  distinct()
# saving the final output write_excel_csv(Epi25OnsetFinal, file = "Epi25OnsetFinal.csv")
```

Figure C.2: R script converting `time since', 'date when', and 'age'into a single common measure of age of onset.

```r
# 1 seizure per time period ----Reading in seizure frequency table

SF <- read_delim("SF_output.csv", delim = ":", escape_double = FALSE, na = "null", trim_ws = TRUE)
SF$DATEREC <-as.Date(SF$DATEREC, "%d/%m/%Y") # to date frm.

# Creating a single column "NumS" from number of seizures (NofS) and Upper number of seizures (UNofS) , ignoring lower no of seizures as no values without the upper range

SF <- SF %>%  mutate(NumS = case_when(!is.na(NofS) ~ NofS,
  is.na(NofS) & UNofS>0 ~ UNofS ))
  SF$NumS <- as.numeric(SF$NumS)  Changing to a numeric value
# Creating a single column "NumTP" from number of time periods (NofTP) and  lower number of  time periods (LNofTP), ignoring
# upper number of time periods as we want the most frequent seizures from the final calculation
SF <- SF %>% mutate(NumTP = case_when(
  NofTP>0 & !is.na(NofTP) ~ NofTP,
  is.na(NofTP) & LNofTP>0 ~ LNofTP ))
  SF$NumTP <- as.numeric(SF$NumTP)  # to numeric value

# 3 Number of Seizures per Time Period ---- Adding a column SperTP which is the result of calculating the number of seizures per # stated time period
SF <- SF %>% mutate(SperTP = NumS/NumTP, .after = "NumTP")

# 4 Seizure frequency per day ---- Creating a column with all time periods as days, we are trying to compare seizure frequency per different time periods, #days would give a common denominator, but it could be done in weeks or months
SF <- SF %>% mutate(Days = case_when(
TP == "Day" ~ 1, TP == "Week" ~ 7,
TP == "Month" ~ 30, TP == "Year" ~ 365))

# Previously calculated number of seizures per time period converted into days
SF <- SF %>% mutate(DailyRate = SperTP/Days)
# 5 EventDate based on separate DD,MD, and YD columns ----
# Converting numerical values given as days (DD) and months (MD) to two figure formats to create full dates from the partial ones
SF <- SF %>% mutate(DD = ifelse(!is.na(DD) & as.numeric(DD) < 10,paste("0", DD, sep = "") , DD))
SF <- SF %>% mutate(MD = ifelse(!is.na(MD) & as.numeric(MD) < 10,paste("0", MD, sep = "") , MD)) ))
SF <- SF %>% mutate(EventDate = ifelse(!is.na(YD) & !is.na(MD) & !is.na(DD), paste(DD, MD, YD, sep = "/"), ifelse(!is.na(YD) & !is.na(MD) & is.na(DD), paste("01", MD, YD, sep = "/"),
ifelse(!is.na(YD) & is.na(MD) & is.na(DD), paste("01","01", YD, sep = "/"),
ifelse(is.na(YD) & !is.na(MD) & is.na(DD), paste("01", MD, ifelse(as.numeric(MD) <  as.numeric(format(DATEREC, "%m")), format(DATEREC, "%Y"),as.character(as.numeric(format(DATEREC, "%Y"))-1)), sep = "/"),
# Get year of or year before (if month mentioned is after month of DATEREC)
ifelse(is.na(YD) & !is.na(MD) & !is.na(DD), paste(DD, MD, ifelse(as.numeric(MD) <  as.numeric(format(DATEREC, "%m")),
# As above but with day also (in above set the 1st of month)
format(DATEREC, "%Y"), as.character(as.numeric(format(DATEREC, "%Y")) -1)), sep = "/"), NA)))), .after = "DATEREC")
```

Figure C.3: R script processing seizure frequency output from ExECT into a dated record of seizure scores.

```r
# When Point in Time is last month, we can take 30 days away from DATEREC
SF <- SF %>% mutate(EventDate2 = case_when(is.na(YD) & is.na(MD) & is.na(DD) & PinT == "Last_Month" ~ DATEREC - 30, is.na(YD) & is.na(MD) & is.na(DD) & PinT == "Last_Year" ~ DATEREC - 365), .after = "EventDate")
# Converting dates to date format
SF$EventDate <- as.Date(SF$EventDate, "%d/%m/%Y")
SF$EventDate2 <- as.Date(SF$EventDate2, "%d/%m/%Y")
# 6 Previous seizure frequency for specific CUI ---- Ordering data by SYSTEM_ID and date of record first really important as we are using lag
SF <- SF %>% arrange(SF, SYSTEM_ID, DATEREC, CUI)
#  PrevRecCUI finds a date of previous clinic - going back up to 6 rows in records sorted by DATEREC /CUI  for the particular CUI i.e., seizure type
SF <- SF %>%  mutate(PrevRecCUI = as.Date(
ifelse(CUI == lag(CUI) & SYSTEM_ID == lag(SYSTEM_ID) & DATEREC > lag(DATEREC), lag(DATEREC),
ifelse(CUI == lag(CUI, 2) & SYSTEM_ID == lag(SYSTEM_ID, 2) & DATEREC > lag(DATEREC, 2), lag(DATEREC, 2),
ifelse(CUI == lag(CUI, 3) & SYSTEM_ID == lag(SYSTEM_ID, 3) & DATEREC > lag(DATEREC, 3) ,lag(DATEREC, 3),
ifelse(CUI == lag(CUI, 4) & SYSTEM_ID == lag(SYSTEM_ID, 4) & DATEREC > lag(DATEREC, 4), lag(DATEREC, 4),
ifelse(CUI == lag(CUI, 5) & SYSTEM_ID == lag(SYSTEM_ID, 5)& DATEREC > lag(DATEREC, 5), lag(DATEREC, 5),
ifelse(CUI == lag(CUI, 6) & SYSTEM_ID == lag(SYSTEM_ID, 6) & DATEREC > lag(DATEREC, 6), lag(DATEREC, 6), NA))))))),.after = "DATEREC" )

# 7 Difference in days ----Difference in days between DATEREC and event date to calculate number of seizures per day for that period
SF <- SF %>% mutate(DaysDiff=case_when(!is.na(EventDate) ~ DATEREC - EventDate, Event date created from dates
is.na(EventDate) & PinT == "LastClinic" & !is.na(PrevRecCUI) ~ DATEREC - PrevRecCUI, is.na(EventDate) & PinT == "Last_Month" |PinT == "Last_Year" ~ DATEREC - EventDate2)) # Event date based on point in time
SF$DaysDiff <- as.numeric(SF$DaysDiff) # to numeric value
# 8 Seizures since per day ---- Shows seizures per day reported as seizures since  (using dates) as DailyRate2
SF <- SF %>% mutate(DailyRate2 = NumS/DaysDiff )
SF <- SF %>% mutate(DailyRateF = case_when(!is.na(DailyRate) ~ DailyRate,
is.na(DailyRate) ~ DailyRate2),.after = "DailyRate2")
# 9 Seizure free----Shows number of days seizure free as based on 0 seizures for a number of days (calculated field) or 0 seizures in the number of days calculated from event day or last clinic, As seizure per time period (SperTP) will always be 0 for 0 seizures we need to take the original number of TP and multiply by days using functions that look back in time
SF <- SF %>% mutate(SeizureFree = case_when(NumS == 0 & !is.na(Days) ~ NumTP*Days, NumS == 0 & is.na(Days) ~ DaysDiff ))

```

Figure C.3: R script processing seizure frequency output from ExECT into a dated record of seizure scores.

```r
# Frequency which was reported for a specific CUI in previous clinic creates PrevCUIFreq looking back 6 rows

SF <- SF %>%  mutate(PrevCUIFreq = case_when
(SYSTEM_ID == lag(SYSTEM_ID) & CUI == lag(CUI) & DATEREC > lag(DATEREC) & is.na(DailyRateF) ~ lag(DailyRateF),
SYSTEM_ID == lag(SYSTEM_ID, 2) &  CUI == lag(CUI, 2) & DATEREC > lag(DATEREC, 2) & is.na(DailyRateF) ~ lag(DailyRateF, 2), SYSTEM_ID == lag(SYSTEM_ID, 3) &  CUI == lag(CUI, 3) & DATEREC > lag(DATEREC, 3) & is.na(DailyRateF) ~ lag(DailyRateF, 3),  SYSTEM_ID == lag(SYSTEM_ID, 4) &  CUI == lag(CUI, 4) & DATEREC > lag(DATEREC, 4) & is.na(DailyRateF) ~ lag(DailyRateF, 4), SYSTEM_ID == lag(SYSTEM_ID, 5) &  CUI == lag(CUI, 5) & DATEREC > lag(DATEREC, 5) & is.na(DailyRateF) ~ lag(DailyRateF, 5), SYSTEM_ID == lag(SYSTEM_ID, 6) &  CUI == lag(CUI, 6) & DATEREC > lag(DATEREC, 6) & is.na(DailyRateF) ~ lag(DailyRateF, 6)), .after = "DailyRateF" )

# 10 "Same" in FreqChange ---- If seizure frequency is reported as "same"  in FreqChange find a letter that is the most recent to the date of "Same" record and use PrevCUIFreq for that record

SF <- SF %>% mutate(PrevRateCUI_Same6m = case_when(FreqChange == "Same" & DATEREC - PrevRecCUI > 182 ~ PrevCUIFreq))
SF <- SF %>% mutate(DailyRateSF = case_when(!is.na(DailyRateF) ~ DailyRateF,
is.na(DailyRateF) ~ PrevRateCUI_Same6m))

# 11 Adding frequency scores ---- Frequency score* calculated as per day rate with seizure free given priority Using the #calculated columns of daily rate (DailyRateSF final rate and rate for "Same" frequency under FreqChange)

SF <- SF %>% mutate(FreqSeverity = case_when(SeizureFree < 7 ~ 6,
SeizureFree > 6 & SeizureFree < 30 ~ 5,
SeizureFree > 29 & SeizureFree < 182 ~ 4,
SeizureFree > 181 & SeizureFree < 365 ~ 3,
SeizureFree > 364 & SeizureFree < 730 ~ 2,
SeizureFree > 729 ~ 1,
DailyRateSF > 1 ~ 7,
DailyRateSF > 0.2857143 &  DailyRateSF < 2 ~ 6,
DailyRateSF > 0.0657534 &  DailyRateSF < 0.2857144 ~ 5,
DailyRateSF > 0.0109589 &  DailyRateSF < 0.0657535 ~ 4,
DailyRateSF > 0.0054795 &  DailyRateSF < 0.0109590 ~ 3,
DailyRateSF > 0.0027322 &  DailyRateSF < 0.0054796 ~ 2,
DailyRateSF < 0.0027321 ~ 1 ))

# SFSeverity dataset ----Creating seizure frequency severity dataset with the original output and some calculated fields and the final severity score

SFSeverity <- select(SF, -PrevRecCUI, -NumS, -NumTP, -SperTP, -DailyRate, -DailyRate2,-DailyRateF, -PrevCUIFreq, -PrevRateCUI_Same6m)

*Frequency scores used here are based on Fitzgerald MP, et.al.2021
```

Figure C.3: R script processing seizure frequency output from ExECT into a dated record of seizure scores.

```

# creating dose in mg only
Prescriptions <- Epi25_PRESC %>% mutate(Quantity_mg = as.numeric(case_when(UNIT == "mg" ~ DOSE,
UNIT == "g" ~ DOSE*1000),.after = "UNIT" ))

#Total daily dose is Frequency x Dose and is called DailyDose and this can be calculetad
#But there are cases when a dose is expressed as 2 or even 3 instances of a single frequency i.e.
#Frequency is given as 1 more than once for the same ASM , We need to find these cases by matching letter (DOC) START (annotation start) and CUI
#prescriptions should be sorted by "DOC" ,"Start" , CUI first

Prescriptions %>%
arrange(SYSTEM_ID, DATEREC, CUI, START, LETTER)
#Find cases when DOSE is a second  or third dose of the same drug - create AnotherDose column

Prescriptions <- Prescriptions %>% mutate (AnotherDose = case_when(LETTER == lag(LETTER) &
START == lag(START) & START != lag(START, 2) & CUI == lag(CUI) & FREQUENCY == '1' & lag(FREQUENCY) == '1' ~ '2nd',
LETTER == lag(LETTER) & LETTER == lag(LETTER,2) & START == lag(START) & START == lag(START, 2) & CUI == lag(CUI) & CUI == lag(CUI, 2) & FREQUENCY == '1' & lag(FREQUENCY, 2) == '1'~ '3rd',
LETTER == lag(LETTER,2) & START == lag(START, 2) & START != lag(START) & FREQUENCY == '1' & lag(FREQUENCY, 2) == '1' ~ '2nd'), .after = "FREQUENCY")

#Extract values of the 2nd and 3rd dose in separate columns but in one row (for some reason adding up lag column references does not work)
Prescriptions <- Prescriptions %>% mutate(Q2ndDose = case_when(AnotherDose == '2nd' &
START == lag(START) & CUI == lag(CUI) ~ lag(Quantity_mg),
AnotherDose == '2nd' & START == lag(START,2) & CUI == lag(CUI,2) ~ lag(Quantity_mg, 2)), .after = "AnotherDose")
Prescriptions <- Prescriptions %>% mutate(Q3rdDose = case_when(AnotherDose == '2nd' & lead(AnotherDose == '3rd') ~ lead(Quantity_mg)), .after = "Q2ndDose")
# DailyDose calculation ----

Prescriptions <- Prescriptions %>% mutate(DailyDose = case_when(FREQUENCY >1 ~ Quantity_mg*FREQUENCY,
!is.na(Q3rdDose) & AnotherDose == '2nd' ~  Quantity_mg + Q2ndDose + Q3rdDose,
is.na(Q3rdDose) & AnotherDose == '2nd' ~ Quantity_mg + Q2ndDose,
is.na(AnotherDose) & is.na(Q2ndDose) & is.na(Q3rdDose) &
START != lead(START) & START != lead(START, 2)  & FREQUENCY == 1 ~ Quantity_mg,
START == lead( START) & LETTER != lead(LETTER) & CUI != lead(CUI)  & FREQUENCY == 1 ~ Quantity_mg,
is.na(FREQUENCY) & is.na(AnotherDose) & CUI == "C0055891" ~ Quantity_mg), .after = "Q3rdDose" )
```

Figure C.4: R script processing prescription output from ExECT into a dated record of the total daily dose of ASM.

```r


#converting brand names to generic names so some continuation can be seen

Prescriptions <- Prescriptions %>%
mutate(PREF = case_when(NAME == "Epilim" ~ "Sodium Valproate",
NAME == "Epilim Chrono" ~ "Sodium Valproate" ,
CUI == "C0591452" ~ "Sodium Valproate" ,
CUI == "C0037567" ~ "Sodium Valproate" ,
NAME == "Lamotrigine" ~ "Lamotrigine",
NAME == "Lamictal" ~ "Lamotrigine",
CUI == "C0064636"~ "Lamotrigine",
NAME == "Levetiracetam" ~ "Levetiracetam",
NAME == "Keppra" ~ "Levetiracetam",
CUI == "C0377265" ~ "Levetiracetam",
CUI == "C2725260" ~ "Eslicarbazepine",
NAME == "Tegretol" ~ "Carbamazepine",
NAME == "Carbamazepine" ~ "Carbamazepine",
CUI == "C0700087" ~ "Carbamazepine",
CUI == "C377265" ~"Carbamazepine",
NAME == "Phenytoin" ~ "Phenytoin",
NAME == "Topiramate" ~ "Topiramate",
CUI == "C0076829" ~ "Topiramate",
NAME == "Topamax" ~ "Topiramate",
NAME == "Topiramate" ~ "Topiramate",
NAME == "Brivaracetam" ~ "Brivaracetam",
NAME == "Clobazam" ~ "Clobazam",
NAME == "Zonisamide" ~ "Zonisamide",
NAME == "Lacosamide" ~ "Lacosamide" ,
NAME == "Perampanel" ~ "Lacosamide" ,
CUI == "C2698764" ~ "Perampanel",
CUI == "C0009011" ~ "Clonazepam",
CUI == "C0060926" ~ "Gabapentin",
CUI == "C2698764" ~ "Perampanel",
CUI == "C0700016" ~ "Primidone",
CUI == "C2725260" ~ "Eslicarbazepine",
CUI == "C0657912"~ "Pregabalin",
CUI == "C0026056" ~ "Midazolam" ))

#selecting a subset with the Daily Dose
DailyPrescription <- select(Prescriptions,SYSTEM_ID, LETTER, DATEREC, CUI, PREF, DailyDose)
# removing NA values from DailyDose (from the double/triple doses)
DailyPrescriptionFull = na.omit(DailyPrescription)
# removing duplicate records = which gives 962 records
DailyPrescriptionFinal =  unique(DailyPrescriptionFull)

# Creating a df of maximum dose for each drug (CUI) and date, slice keeps the max daily dose, removing any lower doses of the same drug, per group
DailyPrescriptionMax <- DailyPrescriptionFinal %>% group_by(SYSTEM_ID, LETTER, DATEREC, CUI, PREF) %>% slice(which.max(DailyDose))

View(DailyPrescriptionMax)  #This is the final output that can be linked to seizure frequency for example
```

Figure C.5: R script processing prescription output from ExECT into a dated record of the total daily dose of ASM.

Appendix D

NLP within SAIL Databank and Genetic Data Linkage

D.1 Linking Epi25 cohort with the SAIL databank datasets

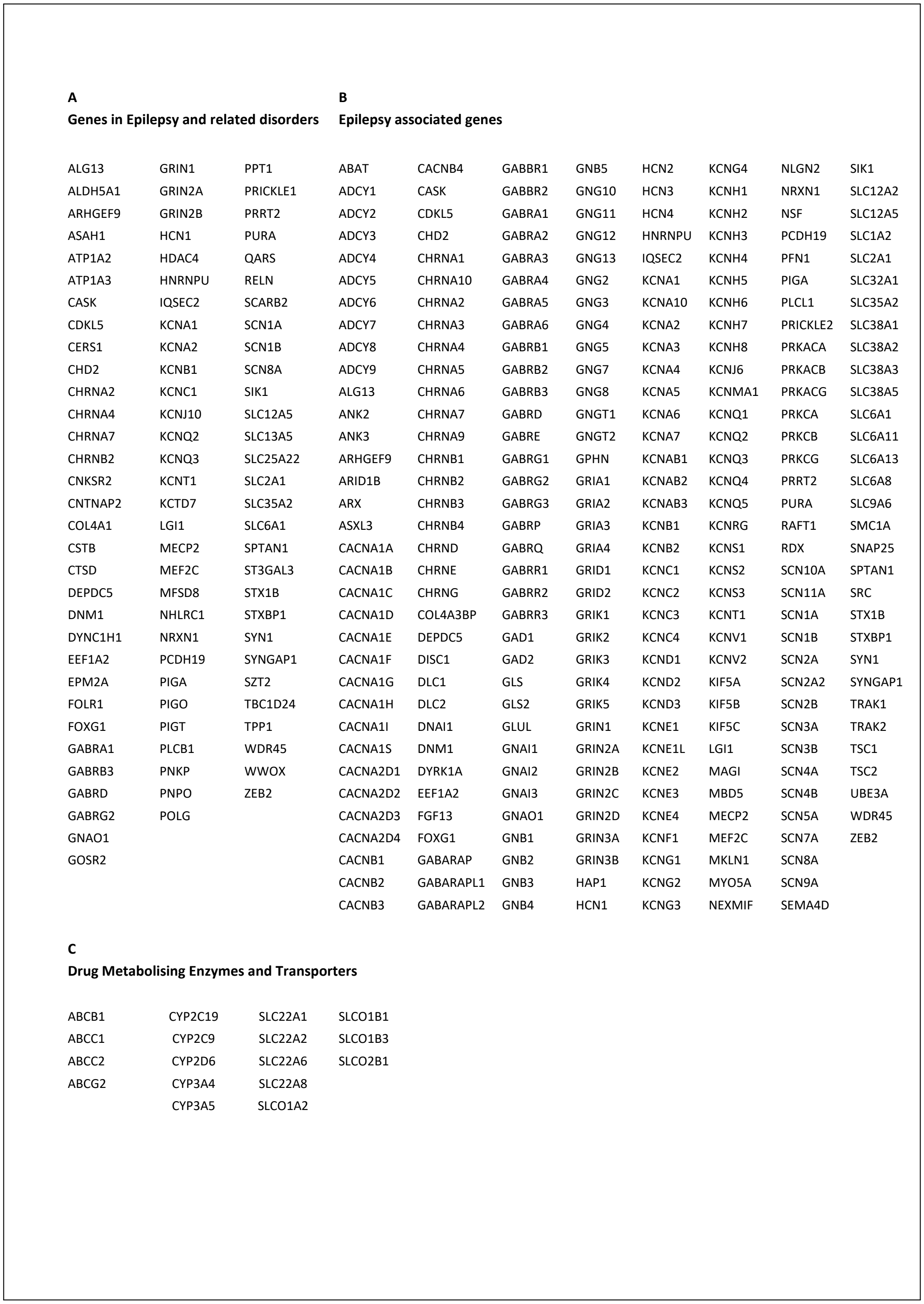
D.2 Epilepsy and drug metabolism and transportation genes

D.3 Filtering for rare and potentially damaging variants

D.4 Residual Variation Intolerance Scores for genes affected by variants identified in the linkage

```

    SELECT
    ea.SYSTEM_ID_PE, -- key field for Epi25 upload based on the biobank number, encrypted
    ea.ALF_PE -- encrypted person ID
    FROM
    SAIL0661V.EPI25_ALF_20190718 ea -- uploded biobank donors (File_1)
    JOIN SAIL0661V.WLGP_PATIENT_ALF_CLEANSED_20200701 gp ON -- SAIL GP registrations
    ea.ALF_PE = gp.ALF_PE -- linking
    GROUP BY
    ea.SYSTEM_ID_PE,
    ea.ALF_PE
    ORDER BY
    ea.SYSTEM_ID_PE,
    ea.ALF_PE;

-------------------------------------------------
    CREATE TEMP TABLE SAIL0661V.Epi25_GP_ASDS AS -- creating temporary table for ASM
    SELECT * FROM
    (SELECT a.* FROM SAIL0661V.Epi25_ALF_SEQ_GP_REG_20200701 a -- Epi25 sequenced cohort linked to GP records
    INNER JOIN SAILV0661.WLGP_PATIENT_ALF_CLEANSED_20200701 b  -- GP ALF (patient registration table)
    ON a.ALF_PE = b.ALF_PE -- linking on ALF
    INNER JOIN SAIL0661V.WLGP_GP_EVENT_CLEANSED_20200701 c -- joining GP events table
    ON b.LOCAL_NUM_PE = c.LOCAL_NUM_PE  AND  b.PRAC_CD_PE = c.PRAC_CD_PE  -- linking by local number identfier, unique number generated during the audit+ extract process for an individual used with the encrypted Practice code to link patients to events
    WHERE EVENT_CD LIKE 'dn%') -- READ code for ASM dn... so dn% captures all ASM
    and EVENT_CD is prescription.
-----------------------------------------
  -- Creating temporary table of individuals who had a record of unscheduled hospital admission with a diagnosis of epilepsy while on ASM treatment
  CREATE TEMP TABLE SAILw0661V.Epi25_SEQ_G40 AS
  SELECT COUNT(*) AS count,a.ALF_PE FROM
  (SELECT e.ALF_PE, e.ADMIS_DT, e.ADMIS_SPEC_CD, e.SPELL_NUM_PE, e.SPELL_DUR, d."First_Event_DT" FROM -- identifying the commencement date of ASM from GP events derived table of sequenced individuals on ASM
  (SELECT ps.ALF_PE, ps.ADMIS_DT, ps.ADMIS_SPEC_CD, ps.SPELL_NUM_PE, ps.SPELL_DUR
  FROM SAIL0661V.PEDW_SPELL_20200901 ps -- linking PEDW admission to ASM from GP records
  JOIN SAILw0661V.Epi25_ALF_SEQ_GP_REG_20200701 gp
  ON ps.ALF_PE = gp.ALF_PE
  where ps.admis_mthd_cd BETWEEN '21' AND '29')e -- admission method for unscheduled admission is 21 and 29
  JOIN SAILw0661V.Epi25_AED_FIRST_EVENT_DT d -- firts ASM prescription
  ON d.ALF_PE = e.ALF_PE
  WHERE e.ADMIS_DT > d."First_Event_DT")a --  date of admission has to be after the date of the first prescription
  JOIN SAIL0661V.PEDW_DIAG_20200901 pd
  ON pd.SPELL_NUM_PE = a.SPELL_NUM_PE
  WHERE pd.DIAG_CD_123  = 'G40' -- Epilepsy ICD-10 diagnosis is G40
  AND pd.DIAG_NUM = 1 -- Epilepsy has to be the primary reason for admission
  GROUP BY a.ALF_PE
  ORDER BY a.ALF_PE;

```

Figure D.1: *Linking Epi25 individuals to GP registrations within the SAIL Databank, GP ASM prescription records, and hospital unscheduled admissions.*



A

Genes in Epilepsy and related disorders

ALG13
ALDH5A1
ARHGEF9
ASAH1
ATP1A2
ATP1A3
CASK
CDKL5
CERS1
CHD2
CHRNA2
CHRNA4
CHRNA7
CHRNB2
CNKSR2
CNTNAP2
COL4A1
CSTB
CTSD
DEPDC5
DNM1
DYNC1H1
EEF1A2
EPM2A
FOLR1
FOXG1
GABRA1
GABRB3
GABRD
GABRG2
GNAO1
GOSR2

GRIN1
GRIN2A
GRIN2B
HCN1
HDAC4
HNRNPU
IQSEC2
KCNA1
KCNA2
KCNB1
KCNC1
KCNJ10
KCNQ2
KCNQ3
KCNT1
KCTD7
LGI1
MECP2
MEF2C
MFSD8
NHLRC1
NRXN1
PCDH19
PIGA
PIGO
PIGT
PLCB1
PNKP
PNPO
POLG

PPT1
PRICKLE1
PRRT2
PURA
QARS
RELN
SCARB2
SCN1A
SCN1B
SCN8A
SIK1
SLC12A5
SLC13A5
SLC25A22
SLC2A1
SLC35A2
SLC6A1
SPTAN1
ST3GAL3
STX1B
STXBP1
SYN1
SYNGAP1
SZT2
TBC1D24
TPP1
WDR45
WWOX
ZEB2

B

Epilepsy associated genes

ABAT
ADCY1
ADCY2
ADCY3
ADCY4
ADCY5
ADCY6
ADCY7
ADCY8
ADCY9
ALG13
ANK2
ANK3
ARHGEF9
ARID1B
ARX
ASXL3
CACNA1A
CACNA1B
CACNA1C
CACNA1D
CACNA1E
CACNA1F
CACNA1G
CACNA1H
CACNA1I
CACNA1S
CACNA2D1
CACNA2D2
CACNA2D3
CACNA2D4
CACNB1
CACNB2
CACNB3

CACNB4
CASK
CDKL5
CHD2
CHRNA1
CHRNA10
CHRNA2
CHRNA3
CHRNA4
CHRNA5
CHRNA6
CHRNA7
CHRNA9
CHRNB1
CHRNB2
CHRNB3
CHRNB4
CHRND
CHRNE
CHRNG
COL4A3BP
DEPDC5
DISC1
DLC1
DLC2
DNAI1
DNM1
DYRK1A
EEF1A2
FGF13
FOXG1
GABARAP
GABARAPL1
GABARAPL2

GABBR1
GABBR2
GABRA1
GABRA2
GABRA3
GABRA4
GABRA5
GABRA6
GABRB1
GABRB2
GABRB3
GABRD
GABRE
GABRG1
GABRG2
GABRG3
GABRP
GABRQ
GABRR1
GABRR2
GABRR3
GAD1
GAD2
GLS
GLS2
GLUL
GNAI1
GNAI2
GNAI3
GNAO1
GNB1
GNB2
GNB3
GNB4

GNB5
GNG10
GNG11
GNG12
GNG13
GNG2
GNG3
GNG4
GNG5
GNG7
GNG8
GNGT1
GNGT2
GPHN
GRIA1
GRIA2
GRIA3
GRIA4
GRID1
GRID2
GRIK1
GRIK2
GRIK3
GRIK4
GRIK5
GRIN1
GRIN2A
GRIN2B
GRIN2C
GRIN2D
GRIN3A
GRIN3B
HAP1
HCN1

HCN2
HCN3
HCN4
HNRNPU
IQSEC2
KCNA1
KCNA10
KCNA2
KCNA3
KCNA4
KCNA5
KCNA6
KCNA7
KCNAB1
KCNAB2
KCNAB3
KCNB1
KCNB2
KCNC1
KCNC2
KCNC3
KCNC4
KCND1
KCND2
KCND3
KCNE1
KCNE1L
KCNE2
KCNE3
KCNE4
KCNF1
KCNG1
KCNG2
KCNG3

KCNG4
KCNH1
KCNH2
KCNH3
KCNH4
KCNH5
KCNH6
KCNH7
KCNH8
KCNJ6
KCNMA1
KCNQ1
KCNQ2
KCNQ3
KCNQ4
KCNQ5
KCNRG
KCNS1
KCNS2
KCNS3
KCNT1
KCNV1
KCNV2
KIF5A
KIF5B
KIF5C
LGI1
MAGI
MBD5
MECP2
MEF2C
MKLN1
MYO5A
NEXMIF

NLGN2
NRXN1
NSF
PCDH19
PFN1
PIGA
PLCL1
PRICKLE2
PRKACA
PRKACB
PRKACG
PRKCA
PRKCB
PRKCG
PRRT2
PURA
RAFT1
RDX
SCN10A
SCN11A
SCN1A
SCN1B
SCN2A
SCN2A2
SCN2B
SCN3A
SCN3B
SCN4A
SCN4B
SCN5A
SCN7A
SCN8A
SCN9A
SEMA4D

SIK1
SLC12A2
SLC12A5
SLC1A2
SLC2A1
SLC32A1
SLC35A2
SLC38A1
SLC38A2
SLC38A3
SLC38A5
SLC6A1
SLC6A11
SLC6A13
SLC6A8
SLC9A6
SMC1A
SNAP25
SPTAN1
SRC
STX1B
STXBP1
SYN1
SYNGAP1
TRAK1
TRAK2
TSC1
TSC2
UBE3A
WDR45
ZEB2

C

Drug Metabolising Enzymes and Transporters

ABCB1
ABCC1
ABCC2
ABCG2

CYP2C19
CYP2C9
CYP2D6
CYP3A4
CYP3A5

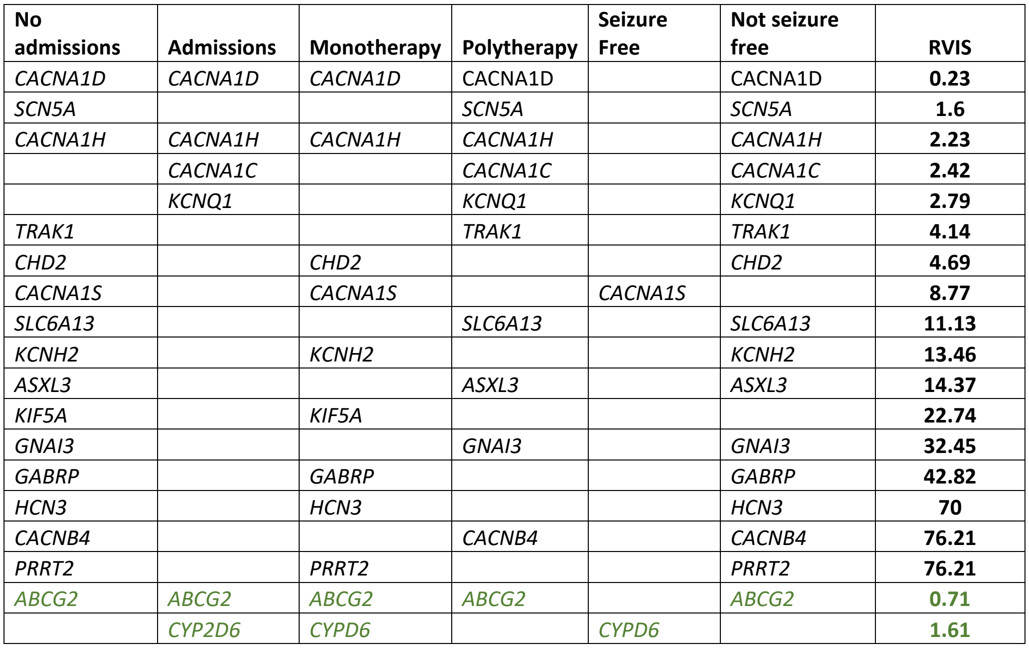
SLC22A1
SLC22A2
SLC22A6
SLC22A8
SLCO1A2

SLCO1B1
SLCO1B3
SLCO2B1

Figure D.2: Genes associated with epilepsy, drug metabolism and transportation used in the gene-based analysis. Source:**A** Epi25 Collaboraive, epilepsy-genes, [1]; **B** Epilepsy associated genes, [2]; **C** Drug metabolising enzymes and transporters. [3]

```
-- creating a temp.table for the filtered data
CREATE TEMPORARY TABLE SAILw0661v.EPI25_AF_001_CADD AS
SELECT * FROM (
SELECT VCF_FILE_PE , CHR,"START", "END","REF",ALT,EXONICFUNC_REFGENE,CADD_PHRED, GENE_REFGENE ,AF
FROM (
SELECT * FROM (
SELECT
VCF_FILE_PE , -- VCF ID which used for linkage
CHR ,"START", "END", "REF", ALT , EXONICFUNC_REFGENE,
0 + CADD_PHRED AS CADD_PHRED, -- converting CADD to numeric
GENE_REFGENE, 0 + AF AS AF  -- Allele frequency (AF) to numeric
FROM
SAIL0661V.EPI25VCF  -- single VCF table for 111 individuals
  WHERE
  AF NOT IN ('-', '.') -- removing blank AF values
  AND CADD_PHRED NOT IN ('-', '.') -- removing blank CADD PHRED values
  LIMIT 100000000 ) -- this stops the process
  WHERE
  AF < 0.001 AND CADD_PHRED >= 15));
-- creating a temporary table of rare and damaging variants
CREATE TEMPORARY TABLE SAILw0661v.EPI25_AF_001_CADD_VAR AS
SELECT CHR || '-' || "START" ||'-'||"END"||'-'|| "REF"||'-' ||ALT  AS VARIANT, GENE_REFGENE , EXONICFUNC_REFGENE, CADD_PHRED, VCF_FILE_PE
FROM SAILw0661v.EPI25_AF_001_CADD
    -- variants not present more than twice in the cohort
CREATE TEMPORARY TABLE SAILw0661v.EPI25_AF_001_CADD_VAR_under3 AS
SELECT "VARIANT","NO_OF_VARIANTS" FROM
(SELECT "VARIANT", COUNT(*) AS "NO_OF_VARIANTS" FROM
(SELECT CHR || '-' || "START" ||'-'||"END"||'-'|| "REF"||'-' ||ALT  AS VARIANT, GENE_REFGENE, EXONICFUNC_REFGENE, CADD_PHRED, VCF_FILE_PE
FROM SAILw0661v.EPI25_AF_001_CADD)
GROUP BY "VARIANT"
ORDER BY "VARIANT")
WHERE "NO_OF_VARIANTS" < 3;

```

Figure D.3: SQL script filtering rare and potentially damaging variants from the annotated whole exome sequenced data table for the study cohort



No admissions	**Admissions**	**Monotherapy**	**Polytherapy**	**Seizure Free**	**Not seizure free**	**RVIS**
CACNA1D	*CACNA1D*	*CACNA1D*	CACNA1D		CACNA1D	**0.23**
SCN5A			*SCN5A*		*SCN5A*	**1.6**
CACNA1H	*CACNA1H*	*CACNA1H*	*CACNA1H*		*CACNA1H*	**2.23**
	CACNA1C		*CACNA1C*		*CACNA1C*	**2.42**
	KCNQ1		*KCNQ1*		*KCNQ1*	**2.79**
TRAK1			*TRAK1*		*TRAK1*	**4.14**
CHD2		*CHD2*			*CHD2*	**4.69**
CACNA1S		*CACNA1S*		*CACNA1S*		**8.77**
SLC6A13			*SLC6A13*		*SLC6A13*	**11.13**
KCNH2		*KCNH2*			*KCNH2*	**13.46**
ASXL3			*ASXL3*		*ASXL3*	**14.37**
KIF5A		*KIF5A*				**22.74**
GNAI3			*GNAI3*		*GNAI3*	**32.45**
GABRP		*GABRP*			*GABRP*	**42.82**
HCN3		*HCN3*			*HCN3*	**70**
CACNB4			*CACNB4*		*CACNB4*	**76.21**
PRRT2		*PRRT2*			*PRRT2*	**76.21**
ABCG2	*ABCG2*	*ABCG2*	*ABCG2*		*ABCG2*	**0.71**
	CYP2D6	*CYPD6*		*CYPD6*		**1.61**

Figure D.4: RVIS for genes identified following filtering for rare and potentially damaging variants and not occurring more than twice in the study cohort in different groups of individuals defined by unscheduled hospital admissions (No admissions versus Admissions), ASM (monotherapy versus polytherapy), and seizure frequency (seizure freedom for > 1 year versus at least 1 seizure per year. Epilepsy associated genes (black font), drug metabolism and transportation (green font))

Bibliography

[1] Epi25 Collaborative. Epi25 Colaborative; 2014. Available from: `http://epi-25.org/genes-in-epilepsy`.

[2] Wang J, Lin ZJ, Liu L, Xu HQ, Shi YW, Yi YH, et al. Epilepsy-associated genes. Seizure. 2017;44:11-20.

[3] Ahmed S, Zhou Z, Zhou J, Chen SQ. Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine; 2016.

[4] Fisher RS, Van Emde Boas W, Blume W, Elger C, Genton P, Lee P, et al. Response: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE) [4]. Epilepsia. 2005;46(10):1701-2.

[5] Fisher RS, Van Emde Boas W, Blume W, Elger C, Genton P, Lee P, et al. Epileptic Seizures and Epilepsy: Definitions Proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia. 2005;46(10):1701-2.

[6] Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE Official Report: A practical clinical definition of epilepsy. Epilepsia. 2014;55(4):475-82.

[7] Giuseppe Capovilla, Anne T Berg, J Helen Cross, Solomon L Moshe, Federico Vigevano, Peter Wolf GA. Conceptual dichotomies in classifyingepilepsies: Partial versus generalized andidiopathic versus symptomatic(April 18–20, 2008, Monreale, Italy). Epilepsia Open. 2009;50(5):1645-9.

[8] Fisher RS, Cross JH, French JA, Higurashi N, Hirsch E, Jansen FE, et al. Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017 apr;58(4):522-30.

[9] Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, Van Emde Boas W, et al. Revised terminology and concepts for organization of seizures and epilepsies: Report of the ILAE Commission on Classification and Terminology, 2005-2009. Epilepsia. 2010;51(4):676-85.

[10] Falco-Walter JJ, Scheffer IE, Fisher RS. The new definition and classification of seizures and epilepsy. Epilepsy Research. 2018;139(July 2017):73-9.

[11] Allen Hauser W, Annegers JF. Descriptive epidemiology of epilepsy: Contributions of population-based studies from rochester, minnesota. Mayo Clinic Proceedings. 1996;71(6):576-86. Available from: `http://dx.doi.org/10.4065/71.6.576`.

[12] Guilhoto LM. Absence epilepsy: Continuum of clinical presentation and epigenetics? Seizure. 2017;44:53-7. Available from: `http://dx.doi.org/10.1016/j.seizure.2016.11.031`.

[13] Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017;58(4):512-21.

[14] Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. Epilepsia. 2017 apr;58(4):512-21.

[15] Nguyen DK, Mbacfou MT, Nguyen DB, Lassonde M. Prevalence of nonlesional focal epilepsy in an adult epilepsy clinic. Canadian Journal of Neurological Sciences. 2013;40(2):198-202.

[16] nabil A. Panayiotopoulos syndrome - YouTube. 2017;324(May):1228-9. Available from: `https://www.youtube.com/watch?v=32uyBnq9Hck`.

[17] Villanueva V, Serratosa JM. Temporal lobe epilepsy: Clinical semiology and age at onset. Epileptic Disorders. 2005;7(2):83-90.

[18] Spooner CG, Berkovic SF, Mitchell LA, Wrennall JA, Harvey AS. New-onset temporal lobe epilepsy in children. Neurology. 2006 dec;67(12):2147 LP 2153. Available from: `http://n.neurology.org/content/67/12/2147.abstract`.

[19] Hirsch E, French J, Scheffer IE, Bogacz A, Alsaadi T, Sperling MR, et al. ILAE definition of the Idiopathic Generalized Epilepsy Syndromes: Position statement by the ILAE Task Force on Nosology and Definitions. Epilepsia. 2022;63(6):1475-99.

[20] Specchio N, Wirrell EC, Scheffer IE, Nabbout R, Riney K, Samia P, et al. International League Against Epilepsy classification and definition of epilepsy syndromes with onset in childhood: Position paper by the ILAE Task Force on Nosology and Definitions. Epilepsia. 2022;63(6):1398-442.

[21] Specchio N, Curatolo P. Developmental and epileptic encephalopathies: What we do and do not know. Brain. 2021;144(1):32-43.

[22] Zack M, Kobau R. Prevalence and incidence of epilepsy: A systematic review and meta-analysis of international studies. Neurology. 2017;89(6):641.

[23] Forsgren L, Beghi E, Õun A, Sillanpää M. The epidemiology of epilepsy in Europe - A systematic review. European Journal of Neurology. 2005;12(4):245-53.

[24] Kotsopoulos IAW, Van Merode T, Kessels FGH, De Krom MCTFM, Knottnerus JA. Systematic review and meta-analysis of incidence studies of epilepsy and unprovoked seizures. Epilepsia. 2002;43(11):1402-9.

[25] Pickrell WO, Lacey AS, Bodger OG, Demmler JC, Thomas RH, Lyons RA, et al. Epilepsy and deprivation, a data linkage study. Epilepsia. 2015;56(4):585-91.

[26] Beghi E, Hesdorffer D. Prevalence of epilepsy - An unknown quantity. Epilepsia. 2014;55(7):963-7.

[27] Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR. Estimation of the burden of active and life-time epilepsy: A meta-analytic approach. Epilepsia. 2010;51(5):883-90.

[28] Morgan CLI, Ahmed Z, Kerr MP. Social deprivation and prevalence of epilepsy and associated health usage. Journal of Neurology Neurosurgery and Psychiatry. 2000;69(1):13-7.

[29] Li X, Sundquist J, Sundquist K. Socioeconomic and occupational risk factors for epilepsy: A nationwide epidemiological study in Sweden. Seizure. 2008;17(3):254-60.

[30] Hesdorffer DC, Tian H, Anand K, Hauser WA, Ludvigsson P, Olafsson E, et al. Socioeconomic status is a risk factor for epilepsy in icelandic adults but not in children. Epilepsia. 2005;46(8):1297-303.

[31] Steer S, Pickrell WO, Kerr MP, Thomas RH. Epilepsy prevalence and socioeconomic deprivation in England. Epilepsia. 2014;55(10):1634-41.

[32] Heaney DC, MacDonald BK, Everitt A, Stevenson S, Leonardi GS, Wilkinson P, et al. Socioeconomic variation in incidence of epilepsy: Prospective community based study in south east England. British Medical Journal. 2002;325(7371):1013-6.

[33] Maloney EM, Corcoran P, Costello DJ, O'Reilly ÉJ. Association between social deprivation and incidence of first seizures and epilepsy: A prospective population-based cohort. Epilepsia. 2022;63(8):2108-19.

[34] Allen Hauser W, Annegers JF. Descriptive epidemiology of epilepsy: Contributions of population-based studies from rochester, minnesota. Mayo Clinic Proceedings. 1996;71(6):576-86. Available from: `http://dx.doi.org/10.4065/71.6.576`.

[35] DAOUD A S BM BATIEHA A, H ES. Risk factors for childhood epilepsy: a case-control study from Irbid, Jordan. Seizure. 2003;1311(03):171-4.

[36] Liu S, Yu W, Lü Y. The causes of new-onset epilepsy and seizures in the elderly. Neuropsychiatric Disease and Treatment. 2016;12:1425-34.

[37] Picot MC, Baldy-Moulinier M, Daurès JP, Dujols P, Crespel A. The prevalence of epilepsy and pharmacoresistant epilepsy in adults: A population-based study in a Western European country. Epilepsia. 2008;49(7):1230-8.

[38] Gastaut H, Gastaut JL, Silva GEGe, Sanchez GRF. Relative Frequency of Different Types of Epilepsy: A Study Employing the Classification of the International League Against Epilepsy. Epilepsia. 1975 sep;16(3):457-61. Available from: `https://doi.org/10.1111/j.1528-1157.1975.tb06073.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1528-1157.1975.tb06073.x`.

[39] Ellis CA, Petrovski S, Berkovic SF. Epilepsy genetics: clinical impacts and biological insights. The Lancet Neurology. 2020;19(1):93-100.

[40] Grinton BE, Heron SE, Pelekanos JT, Zuberi SM, Kivity S, Afawi Z, et al. Familial neonatal seizures in 36 families: Clinical and genetic features correlate with outcome. Epilepsia. 2015;56(7):1071-80.

[41] Dibbens LM, De Vries B, Donatello S, Heron SE, Hodgson BL, Chintawar S, et al. Mutations in DEPDC5 cause familial focal epilepsy with variable foci. Nature Genetics. 2013;45(5):546-51.

[42] Suls A, Jaehn JA, Kecskés A, Weber Y, Weckhuysen S, Craiu DC, et al. De novo loss-of-function mutations in CHD2 cause a fever-sensitive myoclonic epileptic encephalopathy sharing features with dravet syndrome. American Journal of Human Genetics. 2013;93(5):967-75.

[43] Noebels JL. Single-gene determinants of epilepsy comorbidity. Cold Spring Harbor Perspectives in Medicine. 2015;5(11):1-13.

[44] McTague A, Howell KB, Cross JH, Kurian MA, Scheffer IE. The genetic landscape of the epileptic encephalopathies of infancy and childhood. The Lancet Neurology. 2016;15(3):304-16. Available from: http://dx.doi.org/10.1016/S1474-4422(15)00250-1.

[45] Berkovic SF, Howell RA, Hay DA, Hopper JL. Epilepsies in twins: Genetics of the major epilepsy syndromes. Annals of Neurology. 1998;43(4):435-45.

[46] Mullen SA, Berkovic SF, Lowenstein DH, Kato M, Cross H, Satishchandra P, et al. Genetic generalized epilepsies. Epilepsia. 2018;59(6):1148-53.

[47] Thakran S, Guin D, Singh P, Singh P, Kukal S, Rawat C, et al.. Genetic landscape of common epilepsies: Advancing towards precision in treatment; 2020.

[48] Allen AS, Bellows ST, Berkovic SF, Bridgers J, Burgess R, Cavalleri G, et al. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. The Lancet Neurology. 2017;16(2):135-43. Available from: http://dx.doi.org/10.1016/S1474-4422(16)30359-3.

[49] Feng YCA, Howrigan DP, Abbott LE, Tashman K, Cerrato F, Singh T, et al. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. American Journal of Human Genetics. 2019 aug;105(2):267-82.

[50] Thom M. Review: Hippocampal sclerosis in epilepsy: A neuropathology review. Neuropathology and Applied Neurobiology. 2014;40(5):520-43.

[51] Ferlazzo E, Gasparini S, Beghi E, Sueri C, Russo E, Leo A, et al. Epilepsy in cerebrovascular diseases: Review of experimental and clinical data with meta-analysis of risk factors. Epilepsia. 2016;57(8):1205-14.

[52] Barlow KM. Traumatic brain injury. Handbook of Clinical Neurology. 2013;112:891-904.

[53] Vecht CJ, Kerkhof M, Duran-Pena A. Seizure Prognosis in Brain Tumors: New Insights and Evidence-Based Management. The Oncologist. 2014;19(7):751-9.

[54] Almannai M, Al Mahmoud RA, Mekki M, El-Hattab AW. Metabolic Seizures. Frontiers in Neurology. 2021;12(July).

[55] Vezzani A, Fujinami RS, White HS, Preux PM, Blümcke I, Sander JW, et al. Infections, inflammation and epilepsy HHS Public Access. vol. 131; 2016. Available from: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4867498/pdf/nihms784188.pdf`.

[56] Annegers JF, Hauser WA, Beghi E, Nicolosi A, Kurland LT. The risk of unprovoked seizures after encephalitis and meningitis. Neurology. 1988 sep;38(9):1407 LP 1407. Available from: `http://n.neurology.org/content/38/9/1407.abstract`.

[57] Jang Y, Kim DW, Yang KI, Byun JI, Seo JG, No YJ, et al. Clinical approach to autoimmune epilepsy. Journal of Clinical Neurology (Korea). 2020;16(4):519-29.

[58] Giussani G, Bianchi E, Beretta S, Carone D, DiFrancesco JC, Stabile A, et al. Comorbidities in patients with epilepsy: Frequency, mechanisms and effects on long-term outcome. Epilepsia. 2021;62(10):2395-404.

[59] Conrad J, Pawlowski M, Dogan M, Kovac S, Ritter MA, Evers S. Seizures after cerebrovascular events: Risk factors and clinical features. Seizure. 2013;22(4):275-82.

[60] Kelley BJ, Rodriguez M. Seizures in Patients with Multiple Sclerosis. CNS Drugs. 2009 oct;23(10):805-15. Available from: `http://link.springer.com/10.2165/11310900-000000000-00000`.

[61] Finelli PF, Cardi JK. Seizure as a cause of fracture. Neurology. 1989 jun;39(6):858-60.

[62] Ding J, Li X, Tian H, Wang L, Guo B, Wang Y, et al. SCN1A Mutation—Beyond Dravet Syndrome: A Systematic Review and Narrative Synthesis. Frontiers in Neurology. 2021;12(December):1-12.

[63] Dafoulas GE, Toulis KA, Mccorry D, Kumarendran B, Thomas GN, Willis BH, et al. Type 1 diabetes mellitus and risk of incident epilepsy: a population-based, open-cohort study. Diabetologia. 2017;60(2):258-61. Available from: `http://dx.doi.org/10.1007/s00125-016-4142-x`.

[64] Tellez-Zenteno JF, Patten SB, Jetté N, Williams J, Wiebe S. Psychiatric comorbidity in epilepsy: A population-based analysis. Epilepsia. 2007;48(12):2336-44.

[65] Téllez-Zenteno JF, Matijevic S, Wiebe S. Somatic comorbidity of epilepsy in the general population in Canada. Epilepsia. 2005;46(12):1955-62.

[66] Amatniek JC, Hauser WA, DelCastillo-Castaneda C, Jacobs DM, Marder K, Bell K, et al. Incidence and predictors of seizures in patients with Alzheimer's disease. Epilepsia. 2006;47(5):867-72.

[67] Noebels J. A perfect storm: Converging paths of epilepsy and Alzheimer's dementia intersect in the hippocampal formation. Epilepsia. 2011;52(SUPPL. 1):39-46.

[68] Fritzsche K, Baumann K, Götz-Trabert K, Schulze-Bonhage A. Dissociative seizures: A challenge for neurologists and psychotherapists. Deutsches Arzteblatt International. 2013;110(15):263-8.

[69] Francis P, Baker GA. Non-epileptic attack disorder (NEAD): A comprehensive review. Seizure. 1999;8(1):53-61.

[70] Watila MM, Balarabe SA, Ojo O, Keezer MR, Sander JW. Overall and cause-specific premature mortality in epilepsy: A systematic review. Epilepsy & behavior : E&B. 2018 oct;87:213-25.

[71] Keezer MR, Bell GS, Neligan A, Novy J, Sander JW. Cause of death and predictors of mortality in a community-based cohort of people with epilepsy. Neurology. 2016 feb;86(8):704 LP 712. Available from: `http://n.neurology.org/content/86/8/704.abstract`.

[72] Harden C, Tomson T, Gloss D, Buchhalter J, Cross JH, Donner E, et al. Practice Guideline Summary: Sudden Unexpected Death in Epilepsy Incidence Rates and Risk Factors: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology and the American Epilepsy Society. Epilepsy Currents. 2017 may;17(3):180-7. Available from: http://journals.sagepub.com/doi/10.5698/1535-7511.17.3.180.

[73] Chen S, Joodi G, Devinsky O, Sadaf MI, Pursell IW, Simpson Jr RJ. Under-reporting of sudden unexpected death in epilepsy. Epileptic Disorders. 2018;20(4):270-8. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1684/epd.2018.0979.

[74] Sveinsson O, Andersson T, Carlsson S, Tomson T. The incidence of SUDEP. Neurology. 2017 jul;89(2):170 LP 177. Available from: http://n.neurology.org/content/89/2/170.abstract.

[75] Löscher W, Klein P. The Pharmacology and Clinical Efficacy of Antiseizure Medications: From Bromide Salts to Cenobamate and Beyond. CNS Drugs. 2021;35(9):935-63. Available from: https://doi.org/10.1007/s40263-021-00827-8.

[76] Schmidt D, Schachter SC. Drug treatment of epilepsy in adults. BMJ (Online). 2014;348.

[77] Baker GA, Jacoby A, Buck D, Stalgis C, Monnet D. Quality of life of people with epilepsy: A European study. Epilepsia. 1997;38(3):353-62.

[78] Mutanana N, Tsvere M, Chiweshe MK. General side effects and challenges associated with anti-epilepsy medication: A review of related literature. African Journal of Primary Health Care and Family Medicine. 2020;12(1):1-5.

[79] Alsfouk BAA, Brodie MJ, Walters M, Kwan P, Chen Z. Tolerability of Antiseizure Medications in Individuals with Newly Diagnosed Epilepsy. JAMA Neurology. 2020;77(5):574-81.

[80] Marson AG, Al-Kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, et al. The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. Lancet. 2007;369(9566):1016-26.

[81] Marson AG, Al-kharusi AM, Alwaidh M, Appleton R, Baker GA, Chadwick DW, et al. 1-s2.0-S0140673607604607-main-2.pdf. 2007;369.

[82] Lee-Lane E, Torabi F, Lacey A, Fonferko-Shadrach B, Harris D, Akbari A, et al. Epilepsy, antiepileptic drugs, and the risk of major cardiovascular events. Epilepsia. 2021;62(7):1604-16.

[83] Lacey AS, Pickrell WO, Thomas RH, Kerr MP, White CP, Rees MI. Educational attainment of children born to mothers with epilepsy. Journal of Neurology, Neurosurgery &amp; Psychiatry. 2018 jul;89(7):736 LP 740. Available from: `http://jnnp.bmj.com/content/89/7/736.abstract`.

[84] Fonseca-Barriendos D, Frías-Soria CL, Pérez-Pérez D, Gómez-López R, Borroto Escuela DO, Rocha L. Drug-resistant epilepsy: Drug target hypothesis and beyond the receptors. Epilepsia Open. 2022;7(S1):S23-33.

[85] Brodie MJ, Barry SJE, Bamagous GA, Norrie JD, Kwan P. Patterns of treatment response in newly diagnosed epilepsy. Neurology. 2012 may;78(20):1548 LP 1554. Available from: `http://n.neurology.org/content/78/20/1548.abstract`.

[86] Chen Z, Brodie MJ, Liew D, Kwan P. Treatment outcomes in patients with newly diagnosed epilepsy treated with established and new antiepileptic drugs a 30-year longitudinal cohort study. JAMA Neurology. 2018;75(3):279-86.

[87] Hao X, Goldberg D, Kelly K, Stephen L, Kwan P, Brodie MJ. Uncontrolled epilepsy is not necessarily the same as drug-resistant epilepsy: Differences between populations with newly diagnosed epilepsy and chronic epilepsy. Epilepsy and Behavior. 2013;29(1):4-6. Available from: `http://dx.doi.org/10.1016/j.yebeh.2013.06.019`.

[88] Suzuki H, Mikuni N, Ohnishi H, Yokoyama R, Enatsu R, Ochi S. Forgetting to take antiseizure medications is associated with focal to bilateral tonic-clonic seizures, as revealed by a cross-sectional study. PLoS ONE. 2020;15(10 October):1-13. Available from: `http://dx.doi.org/10.1371/journal.pone.0240082`.

[89] Kerr MP. The impact of epilepsy on patients' lives. Acta Neurologica Scandinavica. 2012;126(S194):1-9.

[90] Benson A, O'Toole S, Lambert V, Gallagher P, Shahwan A, Austin JK. The stigma experiences and perceptions of families living with epilepsy: Implications for epilepsy-related communication within and external to the family unit. Patient Education and Counseling. 2016;99(9):1473-81. Available from: `http://dx.doi.org/10.1016/j.pec.2016.06.009`.

[91] Jacoby A, Gorry J, Baker GA. Employers' attitudes to employment of people with epilepsy: Still the same old story? Epilepsia. 2005;46(12):1978-87.

[92] Bishop M, Allen CA. The impact of epilepsy on quality of life: A qualitative analysis. Epilepsy and Behavior. 2003;4(3):226-33.

[93] Baulac M, De Boer H, Elger C, Glynn M, Kälviäinen R, Little A, et al. Epilepsy priorities in Europe: A report of the ILAE-IBE Epilepsy Advocacy Europe Task Force. Epilepsia. 2015;56(11):1687-95.

[94] Nicholls SG, Langan SM, Benchimol EI. Routinely collected data: The importance of high-quality diagnostic coding to research. Cmaj. 2017;189(33):E1054-5.

[95] Poloniecki JD, Atkinson RW, De Leon AP, Anderson HR. Daily time series for cardiovascular hospital admissions and previous day's air pollution in London, UK. Occupational and Environmental Medicine. 1997;54(8):535-40.

[96] Kovats RS, Hajat S, Wilkinson P. Contrasting patterns of mortality and hospital admissions during hot weather and heat waves in Greater London, UK. Occupational and Environmental Medicine. 2004;61(11):893-8.

[97] Bhatnagar P, Wickramasinghe K, Wilkins E, Townsend N. Trends in the epidemiology of cardiovascular disease in the UK. Heart. 2016;102(24):1945-52.

[98] Freeman JA, Hobart JC, Playford ED, Undy B, Thompson AJ. Evaluating neurorehabilitation: Lessons from routine data collection. Journal of Neurology, Neurosurgery and Psychiatry. 2005;76(5):723-8.

[99] Jarman B, Gault S, Alves B, Hider A, Dolan S, Cook A, et al. Explaining differences in English hospital death rates using routinely collected data. British Medical Journal. 1999;318(7197):1515-20.

[100] Nair M, Kurinczuk JJ, Knight M. Establishing a national maternal morbidity outcome indicator in England: A population- based study using routine hospital data. PLoS ONE. 2016;11(4):1-17.

[101] Carter B, Bennett CV, Bethel J, Jones HM, Wang T, Kemp A. Identifying cerebral palsy from routinely-collected data in England and Wales. Clinical Epidemiology. 2019;11:457-68.

[102] Benson T. The history of the Read Codes: The inaugural James Read memorial lecture 2011. Informatics in Primary Care. 2012;19(3):173-82.

[103] Susan E Campbell, Marion K Campbell JMG, Walker AE. Systematic review of discharge coding accuracy. Journal of Public Health. 2012;34(1):138-48.

[104] Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. Journal of Public Health. 2012;34(1):138-48.

[105] Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: A systematic review. Family Practice. 2004;21(4):396-412.

[106] Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: A systematic review. British Journal of General Practice. 2010;60(572):199-206.

[107] Bajaj Y, Crabtree J, Tucker AG. Clinical coding: How accurately is it done? Clinical Governance. 2007;12(3):159-69.

[108] Wang C, Akella R. A Hybrid Approach to Extracting Disorder Mentions from Clinical Notes. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science. 2015 mar;2015:183-7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26306265https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525272/.

[109] Tulloch JSP, Beadsworth MBJ, Vivancos R, Radford AD, Warner JC, Christley RM. GP coding behaviour for non-specific clinical presentations: A pilot study. BJGP Open. 2020;4(3):1-11.

[110] Holman CDA, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: Strategic design, applications and benefits of the WA data linkage system. Australian Health Review. 2008;32(4):766-77.

[111] St Sauver JL, Grossardt BR, Yawn BP, Melton LJ, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: The rochester epidemiology project. American Journal of Epidemiology. 2011;173(9):1059-68.

[112] Irvine K, Hall R, Taylor L. A profile of the Centre for Health Record Linkage. International journal of population data science. 2019;4(2):1142.

[113] Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. Health Information Management Journal. 2008;37(2):60-2. Available from: https://doi.org/10.1177/183335830803700208.

[114] Cnudde P, Rolfson O, Nemes S, Kärrholm J, Rehnberg C, Rogmark C, et al. Linking Swedish health data registers to establish a research database and a shared decision-making tool in hip replacement. BMC Musculoskeletal Disorders. 2016;17(1):1-10. Available from: http://dx.doi.org/10.1186/s12891-016-1262-x.

[115] Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. European Journal of Epidemiology. 2019;34(1):91-9. Available from: `https://doi.org/10.1007/s10654-018-0442-4`.

[116] Blomgren J, Virta LJ. Socioeconomic differences in use of public, occupational and private health care: A register-linkage study of a working-age population in Finland. PLoS ONE. 2020;15(4):1-18. Available from: `http://dx.doi.org/10.1371/journal.pone.0231792`.

[117] Secure Anonymised Information Linkage Databank;. Available from: `https://saildatabank.com/`.

[118] Fonferko-Shadrach B, Lacey AS, White CP, Powell HWR, Sawhney IMS, Lyons RA, et al. Validating epilepsy diagnoses in routinely collected data. Seizure - European Journal of Epilepsy. 2017 nov;52:195-8. Available from: `http://dx.doi.org/10.1016/j.seizure.2017.10.008`.

[119] Mbizvo GK, Bennett KH, Schnier C, Simpson CR, Duncan SE, Chin RFM. The accuracy of using administrative healthcare data to identify epilepsy cases: A systematic review of validation studies. Epilepsia. 2020;61(7):1319-35.

[120] Gorton HC, Webb RT, Carr MJ, DelPozo-Banos M, John A, Ashcroft DM. Risk of unnatural mortality in people with epilepsy. JAMA Neurology. 2018;75(8):929-38.

[121] Pickrell WO, Kerr MP. SUDEP and mortality in epilepsy: The role of routinely collected healthcare data, registries, and health inequalities. Epilepsy and Behavior. 2020;103:106453. Available from: `https://doi.org/10.1016/j.yebeh.2019.106453`.

[122] Mbizvo GK, Schnier C, Simpson CR, Duncan SE, Chin RFM. Validating the accuracy of administrative healthcare data identifying epilepsy in deceased adults: A Scottish data linkage study. Epilepsy Research.

2020;167(June):106462. Available from: https://doi.org/10.1016/j.eplepsyres.2020.106462.

[123] Szpindel A, Myers KA, Ng P, Dorais M, Koclas L, Pigeon N, et al. Epilepsy in children with cerebral palsy: a data linkage study. Developmental Medicine and Child Neurology. 2022;64(2):259-65.

[124] Carson L, Parlatini V, Safa T, Baig B, Shetty H, Phillips-Owen J, et al. The association between early childhood onset epilepsy and attention-deficit hyperactivity disorder (ADHD) in 3237 children and adolescents with Autism Spectrum Disorder (ASD): a historical longitudinal cohort data linkage study. European Child and Adolescent Psychiatry. 2022;(0123456789). Available from: https://doi.org/10.1007/s00787-022-02041-3.

[125] Wotton CJ, Goldacre MJ. Coexistence of schizophrenia and epilepsy: Record-linkage studies. Epilepsia. 2012;53(4).

[126] Allen AN, Seminog OO, Goldacre MJ. Association between multiple sclerosis and epilepsy: Large population-based record-linkage studies. BMC Neurology. 2013;13:2-7.

[127] Schnier C, Duncan S, Wilkinson T, Mbizvo GK, Chin RFM. A nationwide, retrospective, data-linkage, cohort study of epilepsy and incident dementia. Neurology. 2020;95(12):e1686-93. Available from: https://n.neurology.org/content/95/12/e1686.

[128] Hargreaves, Dougal S; Arora, Sandeepa; Viveiro, Carolina; Hale, Daniel R; Ward, Joseph L; Sherlaw-Johnson, Christopher; Viner, Russell M; Dunkley, Colin; Cross JH. Association of quality of paediatric epilepsy care with mortality and unplanned hospital admissions among children and young people with epilepsy in England: a national longitudinal data linkage study. The Lancet Child & Adolescent Health. 2019;3(9):627-35.

[129] Charlton RA, Weil JG, Cunnington MC, Ray S, De Vries CS. Comparing the general practice research database and the UKepilepsy and pregnancy register as tools for postmarketing teratogen surveillance: A nticonvulsants

and the risk of major congenital malformations. Drug Safety. 2011;34(2):157-71.

[130] Pickrell WO, Lacey AS, Bodger OG, Demmler JC, Thomas RH, Lyons RA, et al. Epilepsy and deprivation, a data linkage study. Epilepsia. 2015 apr;56(4):585-91.

[131] Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. npj Digital Medicine. 2019;2(1):1-7. Available from: `http://dx.doi.org/10.1038/s41746-019-0208-8`.

[132] Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, et al. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. Journal of Biomedical Informatics. 2014;51:272-9. Available from: `http://dx.doi.org/10.1016/j.jbi.2014.06.006`.

[133] Rau CS, Kuo PJ, Chien PC, Huang CY, Hsieh HY, Hsieh CH. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. PloS one. 2018 nov;13(11):e0207192-2. Available from: `https://www.ncbi.nlm.nih.gov/pubmed/30412613https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6226171/`.

[134] Fu S, Leung LY, Wang Y, Raulli AO, Kallmes DF, Kinsman KA, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR Medical Informatics. 2019;7(2).

[135] Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: A methodology review. Journal of Biomedical Informatics. 2020;109(August):103526. Available from: `https://doi.org/10.1016/j.jbi.2020.103526`.

[136] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics. 2018 jan;77:34-49. Available from: `https://linkinghub.elsevier.com/retrieve/pii/S1532046417302563`.

[137] Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying Patient Smoking Status from Medical Discharge Records. Journal of the American Medical Informatics Association. 2008;15(1):14-24.

[138] Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo Clinic NLP System for Patient Smoking Status Identification. Journal of the American Medical Informatics Association. 2008;15(1):25-8.

[139] Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: A medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association. 2010;17(1):19-24.

[140] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;17(5):507-13.

[141] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2018;25(3):331-6.

[142] Friedman C. Towards a Comprehensive Medical Language Processing System: Methods and Issues. Journal of the American Medical Informatics Association. 1997;4(SUPPL.):595-9.

[143] Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The {Stanford} {CoreNLP} Natural Language Processing Toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations; 2014. p. 55-60. Available from: http://www.aclweb.org/anthology/P/P14/P14-5010.

[144] Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Computational Biology. 2013;9(2).

[145] Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:931.

[146] Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, et al. Towards a semantic lexicon for clinical natural language processing. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012(4):568-76.

[147] National Library of Medicine (US). UMLS® Reference Manual [Internet]. Bethesda (MD); 2009. Available from: https://www.ncbi.nlm.nih.gov/books.

[148] Pustejovsky J, Stubbs A. Natural Language Annotation for Machine Learning – A guide to Corpus-building for applications; 2013.

[149] Sagheb E, Ramazanian T, Tafti AP, Fu S, Kremers WK, Berry DJ, et al. Use of Natural Language Processing Algorithms to Identify Common Data Elements in Operative Notes for Knee Arthroplasty. Journal of Arthroplasty. 2021;36(3):922-6. Available from: https://doi.org/10.1016/j.arth.2020.09.029.

[150] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011;18(5):552-6.

[151] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics. 2015;58:S20-9.

[152] Fu S, Lopes GS, Pagali SR, Thorsteinsdottir B, Lebrasseur NK, Wen A, et al. Ascertainment of Delirium Status Using Natural Language Processing from Electronic Health Records. Journals of Gerontology - Series A Biological Sciences and Medical Sciences. 2022;77(3):524-30.

[153] Chilman N, Song X, Roberts A, Tolani E, Stewart R, Chui Z, et al. Text mining occupations from the mental health electronic health record: A nat-

ural language processing approach using records from the Clinical Record Interactive Search (CRIS) platform in south London, UK. BMJ Open. 2021;11(3):1-11.

[154] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics. 2009;42(5):950-66. Available from: http://dx.doi.org/10.1016/j.jbi.2008.12.013.

[155] Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. BMC Medical Informatics and Decision Making. 2006;6:1-9.

[156] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics. 2012 oct;45(5):885-92.

[157] Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy. 2013;9(3):330-8. Available from: http://dx.doi.org/10.1016/j.sapharm.2012.04.004.

[158] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37-46.

[159] Dalianis H. Clinical text mining: Secondary use of electronic patient records. Springer International Publishing; 2018.

[160] Helen D, Lacey AS, Mikadze D, Akbari A, Fonferko-Shadrach B, Hollinghurst J, et al. Epilepsy mortality in Wales during COVID-19. Seizure. 2022;94(November 2021):39-42. Available from: https://doi.org/10.1016/j.seizure.2021.11.017.

[161] Cui L, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. EpiDEA: extracting structured epilepsy and seizure information from patient discharge sum-

maries for cohort identification. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:1191-200.

[162] Sullivan R, Yao R, Jarrar R, Buchhalter J, Gonzalez G. Text Classification towards Detecting Misdiagnosis of an Epilepsy Syndrome in a Pediatric Population. AMIA Annual Symposium proceedings AMIA Symposium. 2014;2014:1082-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25954418{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4419916.

[163] Decker BM, Turco A, Xu J, Terman SW, Kosaraju N, Jamil A, et al. Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. Seizure: European Journal of Epilepsy. 2022;101(July):48-51. Available from: https://doi.org/10.1016/j.seizure.2022.07.010.

[164] Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, et al. Extracting seizure frequency from epilepsy clinic notes: A machine reading approach to natural language processing. Journal of the American Medical Informatics Association. 2022;29(5):873-81.

[165] Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: Development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open. 2019 apr;9(4).

[166] Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: Current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3):1-22.

[167] Leu C, Balestrini S, Maher B, Hernández-Hernández L, Gormley P, Hämäläinen E, et al. Genome-wide Polygenic Burden of Rare Deleterious Variants in Sudden Unexpected Death in Epilepsy. EBioMedicine. 2015;2(9):1063-70. Available from: http://dx.doi.org/10.1016/j.ebiom.2015.07.005.

[168] Naimo GD, Guarnaccia M, Sprovieri T, Ungaro C, Conforti FL, Andò S, et al. A systems biology approach for personalized medicine in refractory epilepsy. International Journal of Molecular Sciences. 2019;20(15):1-15.

[169] Glauser TA, Holland K, O'Brien VP, Keddache M, Martin LJ, Clark PO, et al. Pharmacogenetics of antiepileptic drug efficacy in childhood absence epilepsy. Annals of Neurology. 2017;81(3):444-53.

[170] Heavin SB, McCormack M, Wolking S, Slattery L, Walley N, Avbersek A, et al. Genomic and clinical predictors of lacosamide response in refractory epilepsies. Epilepsia Open. 2019;4(4):563-71.

[171] Wolking S, Schulz H, Nies AT, Mccormack M, Schaeffeler E, Auce P, et al. Pharmacoresponse in genetic generalized epilepsy: a genome-wide association study. Pharmacogenomics. 2020;21(5):325-35.

[172] Wolking S, Moreau C, Nies AT, Schaeffeler E, McCormack M, Auce P, et al. Testing association of rare genetic variants with resistance to three common antiseizure medications. Epilepsia. 2020 apr;61(4):657-66. Available from: `https://onlinelibrary.wiley.com/doi/epdf/10.1111/epi.16467`.

[173] Wolking S, Campbell C, Stapleton C, McCormack M, Delanty N, Depondt C, et al. Role of Common Genetic Variants for Drug-Resistance to Specific Anti-Seizure Medications. Frontiers in Pharmacology. 2021;12(June):1-7.

[174] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003;31(13):3812-4.

[175] Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014;46(3):310-5.

[176] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019;47(D1):D886-94.

[177] Team R. RStudio: Integrated Development Environment for R. 2020. Available from: `https://www.rstudio.com/`.

[178] Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:144-53.

[179] Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: Natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. Journal of the American Medical Informatics Association. 2014;21(3):406-13.

[180] Artstein R. Inter-annotator agreement. 2017:297-313.

[181] Teruel M, Cardellino C, Cardellino F, Alemany LA, Villata S. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. LREC 2018 - 11th International Conference on Language Resources and Evaluation. 2019:4061-4.

[182] Fleiss JL. Nominal Scale Among Many Rater. Psychological Bulletin. 1971;76(5):378-82.

[183] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data; 1977. 1.

[184] Viera AJ, Garrett JM. Anthony J. Viera, MD; Joanne M. Garrett, PhD (2005). Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-63. Family Medicine. 2005;37(5):360-3. Available from: `http://www1.cs.columbia.edu/{~}julia/courses/CS6998/Interrater{\_}agreement.Kappa{\_}statistic.pdf`.

[185] Raghavan P, Fosler-Lussier E, Lai AM. Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:1366-74.

[186] Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association. 2005;12(3):296-8.

[187] Trivedi S, Gildersleeve R, Franco S, Kanter AS, Chaudhry A. Evaluation of a Concept Mapping Task Using Named Entity Recognition and Normalization in Unstructured Clinical Text. Journal of Healthcare Informatics Research. 2020;4(4):395-410.

[188] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for {NLP}-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the {E}uropean Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics; 2012. p. 102-7. Available from: `https://aclanthology.org/E12-2021`.

[189] Dobbie S, Strafford H, Pickrell WO, Fonferko-Shadrach B, Jones C, Akbari A, et al. Markup: A Web-Based Annotation Tool Powered by Active Learning. Frontiers in Digital Health. 2021;3(July):1-9.

[190] Cunningham H. Developing Language Processing Components with GATE Version 8. University of Sheffield Department of Computer Science; 2014. Available from: `https://gate.ac.uk/sale/tao/split.html`.

[191] Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009;42(5):839-51.

[192] Chinchor N. MUC-4 evaluation metrics. 4th Message Understanding Conference, MUC 1992 - Proceedings. 1992:22-9.

[193] R Core Team. R: A language and environment for statistical computing.. Vienna, Austria.: R Foundation for Statistical Computing,; 2021.

[194] G Grothendieck. sqldf R package; 2017. Available from: `https://github.com/ggrothendieck/sqldf`.

[195] Müller HW, François R, Henry L, Kirill. dplyr: A Grammar of Data Manipulation; 2022. Available from: `https://github.com/tidyverse/dplyr`.

[196] Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nature Protocols. 2015;10(10).

[197] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

[198] Strachan, T , Goodship, J , Chinnery PF. Genetics and genomics in medicine; 2015. Available from: `http://worldcat.org`.

[199] Carol Deutsch1 Pengse Po1 Erin Delaney1 HGMSKLSLTAKEJPYMH. HHS Public Access. Physiology & behavior. 2017;176(12):139-48.

[200] Lyons RA, Jones KH, John G, Brooks CJ, Verplancke JP, Ford DV, et al. The SAIL databank: Linking multiple health and social care datasets. BMC Medical Informatics and Decision Making. 2009;9(1):1-8.

[201] Jones KH, Heys S, Thompson R, Cross L, Ford D. International Journal of study of the SAIL Databank. 2020;0(August).

[202] Best AV, Akbari A, Bedston S, Lowthian E, Lyons J, Torabi F, et al. Informing better use of Welsh hospital admissions data in the SAIL Databank : A review of the clinical coding completeness and lag in the Patient Episode Database for Wales (PEDW). 2022;10(March 2020).

[203] Patient Episode Database for Wales (PEDW); 2022. Available from: `https://dhcw.nhs.wales/information-services/health-intelligence/pedw-data-online/`.

[204] Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag New York; 2016. Available from: `https://ggplot2.tidyverse.org`.

[205] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 jan;47(D1):D886-94.

[206] Jiang D, Niwa M, Koong AC, Diego S. Stdg. 2016;10(10):48-56.

[207] Axosoft. GitKraken; 2022. Available from: `https://www.gitkraken.com/`.

[208] NCBI. PumMed;. Available from: `https://pubmed.ncbi.nlm.nih.gov/`.

[209] Elsevier. No Title; 2022. Available from: `https://www.sciencedirect.com/`.

[210] Benito van der Zander, Jan Sundermeyer, Daniel Braun TH. TeXstudio;. Available from: `https://www.texstudio.org/`.

[211] Canales L, Menke S, Marchesseau S, D'Agostino A, del Rio-Bermudez C, Taberna M, et al. Assessing the performance of clinical natural language processing systems: Development of an evaluation methodology. JMIR Medical Informatics. 2021;9(7):1-13.

[212] Patterson BW, Jacobsohn GC, Shah MN, Song Y, Maru A, Venkatesh AK, et al. Development and validation of a pragmatic natural language processing approach to identifying falls in older adults in the emergency department. BMC Medical Informatics and Decision Making. 2019;19(1):1-8.

[213] Shorvon S. The concept of symptomatic epilepsy and the complexities of assigning cause in epilepsy. Epilepsy and Behavior. 2014;32:1-8. Available from: `http://dx.doi.org/10.1016/j.yebeh.2013.12.025`.

[214] Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, et al. The SAIL Databank: Building a national architecture for e-health research and evaluation. BMC Health Services Research. 2009;9(June 2014).

[215] Fitzgerald MP, Kaufman MC, Massey SL, Fridinger S, Prelack M, Ellis C, et al. Assessing seizure burden in pediatric epilepsy using an electronic medical record-based tool through a common data element approach CHOP

Pediatric Epilepsy Program Collaborative explained in Acknowledgments section. 2021.

[216] Committee JF. British National Formulary (online) London; 2022. Available from: `http://www.medicinescomplete.com`.

[217] Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nature reviews Genetics. 2007 oct;8(10):749-61.

[218] Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genetics. 2013;9(8).

[219] Dwivedi AK, Mallawaarachchi I, Alvarado LA. Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. Statistics in Medicine. 2017;36(14):2187-205.

[220] Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, et al. The SAIL Databank: Building a national architecture for e-health research and evaluation. BMC Health Services Research. 2009;9.

[221] Bootsma HPR, Ricker L, Diepman L, Gehring J, Hulsman J, Lambrechts D, et al. Long-term effects of levetiracetam and topiramate in clinical practice: A head-to-head comparison. Seizure. 2008;17(1):19-26.

[222] Velupillai S, Dalianis H, Kvist M. Factuality levels of diagnoses in Swedish clinical text. Studies in Health Technology and Informatics. 2011;169:559-63.

[223] Jagannatha A, Liu F, Liu W, Yu H. Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0). Drug Safety. 2019;42(1):99-111. Available from: `https://doi.org/10.1007/s40264-018-0762-z`.

[224] Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF corpus: semantic annotation of clinical text. AMIA Annual

Symposium proceedings / AMIA Symposium AMIA Symposium. 2007:625-9.

[225] Viani N, Botelle R, Kerwin J, Yin L, Patel R, Stewart R, et al. A natural language processing approach for identifying temporal disease onset information from mental healthcare text. Scientific Reports. 2021;11(1):1-12. Available from: https://doi.org/10.1038/s41598-020-80457-0.

[226] Zhao M, Havrilla J, Peng J, Drye M, Fecher M, Guthrie W, et al. Development of a phenotype ontology for autism spectrum disorder by natural language processing on electronic health records. Journal of Neurodevelopmental Disorders. 2022;14(1):1-12. Available from: https://doi.org/10.1186/s11689-022-09442-0.

[227] Jeffrey P Yaeger MD, MPH, Jiahao Lu BS, Jeremiah Jones MA, Ashkan Ertefaie PhD, Kevin Fiscella MD DGP. Derivation of a natural language processing algorithm to identify febrile infants. Journal of Hospital Medicine. 2022;17(1):11-8. Available from: https://shmpublications.onlinelibrary.wiley.com/doi/10.1002/jhm.2732.

[228] Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records : Overview of 2016 CEGS N-GRID shared tasks Track 1. Journal of Biomedical Informatics. 2017;75:S4-S18. Available from: https://doi.org/10.1016/j.jbi.2017.06.011.

[229] Neamatullah I, Douglass MM, Lehman LWH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Medical Informatics and Decision Making. 2008;8:1-17.

[230] U S Department of Health and Human Services Office for Civil Rights. HIPAA Administrative Simplification. U.S. Department of Health and Human Services; 2009. Available from: https://www.hhs.gov/sites/default/files/hipaa-simplification-201303.pdf.

[231] Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across

institutions. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2012;2012:577-86.

[232] Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association. 2012;19(E1):162-9.

[233] Patterson O, Hurdle JF. Document clustering of clinical narratives: a systematic study of clinical sublanguages. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:1099-107.

[234] Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Briefings in Bioinformatics. 2021;22(1):146-63.

[235] Boisen S, Crystal MR, Schwartz R, Stone R, Weischedel R. Annotating resources for information extraction. 2nd International Conference on Language Resources and Evaluation, LREC 2000. 2000:87-90.

[236] Mohammad HA, Sivarajkumar S, Viggiano S, Oniani D, Visweswaran S, Wang Y. Extraction of Sleep Information from Clinical Notes of Alzheimer ' s Disease Patients Using Natural Language Processing. 2022.

[237] Sim J, Wright CC. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy. 2005;85(3):257-68.

[238] Koleck TA, Tatonetti NP, Bakken S, Mitha S, Henderson MM, George M, et al. Identifying symptom information in clinical notes using natural language processing. Nursing Research. 2021;70(3):173-83.

[239] Sykes D, Grivas A, Grover C, Tobin R, Sudlow C, Whiteley W, et al. Comparison of rule-based and neural network models for negation detection in radiology reports. Natural Language Engineering. 2021;27(2):203-24.

[240] Skeppstedt M. Extracting Clinical Findings from Swedish Health Record Text. Stockholm University; 2014.

[241] Jones BE, South BR, Shao Y, Lu CC, Leng J, Sauer BC, et al. Development and Validation of a Natural Language Processing Tool to Identify Patients Treated for Pneumonia across VA Emergency Departments. Applied Clinical Informatics. 2018;9(1):122-8.

[242] Quimbaya AP, Múnera AS, Rivera RAG, Rodríguez JCD, Velandia OMM, Peña AAG, et al. Named Entity Recognition over Electronic Health Records Through a Combined Dictionary-based Approach. Procedia Computer Science. 2016;100:55-61.

[243] Zamrodah Y. No Title No Title No Title. 2016;15(2):1-23.

[244] Dobbie S, Strafford H, Pickrell WO, Fonferko-Shadrach B, Jones C, Akbari A, et al. Markup: A Web-Based Annotation Tool Powered by Active Learning. Frontiers in Digital Health. 2021 jul;3.

[245] Hamid H, Fodeh SJ, Lizama AG, Czlapinski R, Pugh MJ, LaFrance WC, et al. Validating a natural language processing tool to exclude psychogenic nonepileptic seizures in electronic medical record-based epilepsy research. Epilepsy and Behavior. 2013;29(3):578-80. Available from: `http://dx.doi.org/10.1016/j.yebeh.2013.09.025`.

[246] Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association. 2010;17(5):514-8.

[247] Mowery DL, Kawamoto K, Bradshaw R, Kohlmann W, Schiffman JD, Weir C, et al. Determining Onset for Familial Breast and Colorectal Cancer from Family History Comments in the Electronic Health Record. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science. 2019;2019:173-81. Available from: `http://www.ncbi.nlm.nih.gov/pubmed/31258969{\%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6568127`.

[248] Viani N, Kam J, Yin L, Bittar A, Dutta R, Patel R, et al. Temporal information extraction from mental health records to identify duration of untreated psychosis. Journal of Biomedical Semantics. 2020;11(1):1-11.

[249] Chipaux M, Szurhaj W, Vercueil L, Milh M, Villeneuve N, Cances C, et al. Epilepsy diagnostic and treatment needs identified with a collaborative database involving tertiary centers in France. Epilepsia. 2016;57(5):757-69.

[250] Fogarasi A, Jokeit H, Faveret E, Janszky J, Tuxhorn I. The effect of age on seizure semiology in childhood temporal lobe epilepsy. Epilepsia. 2002;43(6):638-43.

[251] Artuso R, Mencarelli MA, Polli R, Sartori S, Ariani F, Pollazzon M, et al. Early-onset seizure variant of Rett syndrome: Definition of the clinical diagnostic criteria. Brain and Development. 2010;32(1):17-24. Available from: http://dx.doi.org/10.1016/j.braindev.2009.02.004.

[252] Jang SS, Kim SY, Kim H, Hwang H, Chae JH, Kim KJ, et al. Diagnostic Yield of Epilepsy Panel Testing in Patients With Seizure Onset Within the First Year of Life. Frontiers in Neurology. 2019;10(September).

[253] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artificial Intelligence in Medicine. 2021 jul;117.

[254] Berg AT, Zelko FA, Levy SR, Testa FM. Age at onset of epilepsy, pharmacoresistance, and cognitive outcomes: A prospective cohort study. Neurology. 2012;79(13):1384-91.

[255] Cramer JA, French J. Quantitative assessment of seizure severity for clinical trials: A review of approaches to seizure components. Epilepsia. 2001;42(1):119-29.

[256] 1000G. The Variant Call Format (VCF) Version 4 . 2 Specification. Online Resource. 2015:1-28. Available from: http://samtools.github.io/hts-specs/VCFv4.2.pdf.

[257] Rodan LH, Spillmann RC, Kurata HT, Lamothe SM, Maghera J, Jamra RA, et al. Phenotypic expansion of CACNA1C-associated disorders to include isolated neurological manifestations. Genetics in Medicine. 2021;23(10):1922-32.

[258] Bozarth X, Dines JN, Cong Q, Mirzaa GM, Foss K, Lawrence Merritt J, et al. Expanding clinical phenotype in CACNA1C related disorders: From neonatal onset severe epileptic encephalopathy to late-onset epilepsy. American Journal of Medical Genetics, Part A. 2018;176(12):2733-9.

[259] Lv N, Qu J, Long H, Zhou L, Cao Y, Long L, et al. Association study between polymorphisms in the CACNA1A, CACNA1C, and CACNA1H genes and drug-resistant epilepsy in the Chinese Han population. Seizure. 2015;30:64-9. Available from: `http://dx.doi.org/10.1016/j.seizure.2015.05.013`.

[260] Goldman AM, Glasscock E, Yoo J, Chen TT, Klassen TL NJ. Arrhythmia in Heart and Brain: KCNQ1 Mutations Link Epilepsy and Sudden Unexplained Death. Science translational medicine. 2009. Available from: `https://pubmed.ncbi.nlm.nih.gov/20368164/`.

[261] Tiron C, Campuzano O, Pérez-Serra A, Mademont I, Coll M, Allegue C, et al. Further evidence of the association between LQT syndrome and epilepsy in a family with KCNQ1 pathogenic variant. Seizure. 2015;25:65-7.

[262] Goddard KAB, Smith KS, Chen C, McMullen C, Johnson C. Biobank recruitment: Motivations for nonparticipation. Biopreservation and Biobanking. 2009 jun;7(2):119-21.

[263] Johnsson L, Helgesson G, Rafnar T, Halldorsdottir I, Chia KS, Eriksson S, et al. Hypothetical and factual willingness to participate in biobank research. European Journal of Human Genetics. 2010;18(11):1261-4.

[264] Perucca P. Genetics of focal epilepsies: What do we know and where are we heading? Epilepsy Currents. 2018;18(6):356-62.

[265] F Dewey, M Murray, J Overton, L Habegger, J Leader, S Fetterolf, C O'Dushlaine, C Van Hout, J Staples, R Metpally, H L Kirchner, S Pen-

dergrass, C Gonzaga-Jauregui, S Balasubramanian, A Lopez, J Penn, S Mukherjee, N Gosalia, A Li, S Bru DC. Distribution and clinical impact of functional variants in 50,726 whole exome sequences from the DiscovEHR study. ASHG. 2016. Available from: http://www.discovehrshare.com/.

[266] Feng YCA, Howrigan DP, Abbott LE, Tashman K, Cerrato F, Singh T, et al. Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. American Journal of Human Genetics. 2019 aug;105(2):267-82.

[267] Leu C, Stevelink R, Smith AW, Goleva SB, Kanai M, Ferguson L, et al. Polygenic burden in focal and generalized epilepsies. Brain. 2019;142(11):3473-81.

[268] Chen J, Zhang J, Liu A, Zhang L, Li H, Zeng Q, et al. CHD2-related epilepsy: novel mutations and new phenotypes. Developmental Medicine and Child Neurology. 2020;62(5):647-53.

[269] Decker BM, Hill CE, Baldassano SN, Khankhanian P. Can antiepileptic efficacy and epilepsy variables be studied from electronic health records? A review of current approaches. Seizure. 2021;85(November 2020):138-44.

[270] Kwan P, Wong V, Ng PW, Lui CHT, Sin NC, Wong KS, et al. Gene-wide tagging study of the association between ABCC2, ABCC5 and ABCG2 genetic polymorphisms and multidrug resistance in epilepsy. Pharmacogenomics. 2011;12(3).

[271] et al Dong Wook K. Lack of association between ABCB1, ABCG2, and ABCC2 genetic polymorphisms and multidrug resistance in partial epilepsy. Epilepsy Research. 2009;84(1):86-90.

[272] Zan X, Yue G, Hao Y, Sima X. A systematic review and meta-analysis of the association of ABCC2/ABCG2 polymorphisms with antiepileptic drug responses in epileptic patients. Epilepsy Research. 2021;175(May):106678. Available from: https://doi.org/10.1016/j.eplepsyres.2021.106678.

[273] Mousavi SF, Hasanpour K, Nazarzadeh M, Adli A, Bazghandi MS, Asadi A, et al. ABCG2, SCN1A and CYP3A5 genes polymorphism and drug-resistant

epilepsy in children: A case-control study. Seizure - European Journal of Epilepsy. 2022 apr;97:58-62. Available from: https://doi.org/10.1016/j.seizure.2022.03.009.