Swansea University

2022

Repeatable and reusable research - Exploring the needs of users for a Data Portal for Disease Phenotyping

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy in Medical and Health Care Studies

Zahra Ahmed Almowil

BSc, MSc

Data Science Building, Medical School

Swansea University

2022

DECLARATION

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

Signed: Zahra Almowil

Date: 17/03/2022

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s). Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed: Zahra Almowil

Date: 17/03/2022

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: Zahra Almowil

Date: 17/03/2022

ACKNOWLEDGEMENT

To begin with, I would like to express my deepest appreciation to my current supervisors, Prof. Sinead and Dr. Jodie, as well as to my previous supervisor, Dr. Shangming, for their continuous encouragement during my PhD, as well as for their patience, motivation, and exceptional knowledge. Their guidance assisted me in the completion of this PhD thesis.

I would like to express my gratitude to my family: my parents, brothers, and sisters, for their support and understanding throughout the time I worked on this thesis. I consider myself fortunate to have you as family.

This thesis would be dedicated to my husband. I appreciate your assistance and support during a time when I was in desperate need.

Finally, I would like to express my gratitude to my lovely son, Hassan, for accompanying me on this great journey and for being my wonderful companion during the difficult days. You've inspired me with the desire to succeed. Hassan, I adore you.

Many thanks to everyone!

Zahra

LIST OF PUBLICATIONS

Almowil Z, Zhou S, Brophy S, Croxall J. Concept Libraries for Repeatable and Reusable Research : Qualitative Study Exploring the Needs of Users Corresponding Author : 9:1–16.

Zahra A, Shang-Ming Z, Sinead B. Concept libraries for automatic electronic health record based phenotyping: A review. Int J Popul Data Sci. 2021;6(1):1–17.

ABSTRACT

Background: Big data research in the field of health sciences is hindered by a lack of agreement on how to identify and define different conditions and their medications. This means that researchers and health professionals often have different phenotype definitions for the same condition. This lack of agreement makes it hard to compare different study findings and hinders the ability to conduct repeatable and reusable research.

Objective: This thesis aims to examine the requirements of various users, such as researchers, clinicians, machine learning experts, and managers, for both new and existing data portals for phenotypes (concept libraries).

Methods: Exploratory sequential mixed methods were used in this thesis to look at which concept libraries are available, how they are used, what their characteristics are, where there are gaps, and what needs to be done in the future from the point of view of the people who use them. This thesis consists of three phases: 1) two qualitative studies, including one-to-one interviews with researchers, clinicians, machine learning experts, and senior research managers in health data science, as well as focus group discussions with researchers working with the Secured Anonymized Information Linkage databank, 2) the creation of an email survey (i.e., the Concept Library Usability Scale), and 3) a quantitative study with researchers, health professionals, and clinicians.

Results: Most of the participants thought that the prototype concept library would be a very helpful resource for conducting repeatable research, but they specified that many requirements are needed before its development. Although all the participants stated that they were aware of some existing concept libraries, most of them expressed negative perceptions about them. The participants mentioned several facilitators that would encourage them to: 1) share their work, such as receiving citations from other researchers; and 2) reuse the work of others, such as saving a lot of time and effort, which they frequently spend on creating new code lists from scratch. They also pointed out several barriers that could inhibit them from: 1) sharing their work, such as concerns about intellectual property (e.g., if they shared their methods before publication, other researchers would use them as their own); and 2) reusing others' work, such as a lack of confidence in the quality and validity of their code lists. Participants suggested some developments that they would like to see happen in order to make research that is done with routine data more reproducible, such as the availability of a drive for more transparency in research methods documentation, such as publishing complete phenotype definitions and clear code lists.

Conclusions: The findings of this thesis indicated that most participants valued a concept library for phenotypes. However, only half of the participants felt that they would contribute by providing definitions for the concept library, and they reported many barriers regarding sharing their work on a publicly accessible platform such as the CALIBER research platform. Analysis of interviews, focus group discussions, and qualitative studies revealed that different users have different requirements, facilitators, barriers, and concerns about concept libraries.

This work was to investigate if we should develop concept libraries in Kuwait to facilitate the development of improved data sharing. However, at the end of this thesis the recommendation is this would be unlikely to be cost effective or highly valued by users and investment in open access research publications may be of more value to the Kuwait research/academic community.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ACSs	Acute Coronary Syndromes
AHPs	Allied Health Professions
ALF	Anonymous Link Field
AoMRC	Academy of Medical Royal Colleges
AS	Ankylosing Spondylitis
AUC	Area Under Curve
BMA	British Medical Association
CAD	Stable Coronary Disease
CCI	Charlson Comorbidity Index
CDS	Central Returns and Commissioning Data Sets
CIMG	Clinical Imaging Management Group
CPRD	Clinical Practice Research Datalink
CSUQ	Computer System Usability Questionnaire
CVDs	Cardiovascular diseases
DAP	Data Access Process
dbGaP	database of Genotypes and Phenotypes
DR	Diabetic Retinopathy
EMRs	Electronic Medical Records
e-RS	electronic Referral Service
FH	Familial Hypercholesterolemia
GP	General Practice
GPAP	Genome-Phenome Analysis Platform
GRU	General Research Usage
HCRW	Health and Care Research Wales
HES	Hospital Episodes Statistics

HPO	Human Phenotype Ontology
ICD	International Classification of Disease
IGRP	Independent Information Governance Review Panel
IHTSDO	International Health Terminology Standards Development Organization
ISAC	Independent Scientific Advisory Committee
JSON	JavaScript Object Notation
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LMW	Low Molecular Weight
LSOA	Lower Layer Super Output Areas
MCHP	Manitoba Centre for Health Policy
MI	Myocardial Infarction
MINAP	Myocardial Ischaemia National Audit Project
ML	Machine learning
NCBI	National Centre for Biotechnology Information
NCRS	NHS Care Record Service
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NICIP	National Interim Clinical Imaging Procedure
NIH	National Institutes of Health
NL	Newfoundland and Labrador
NLP	Natural Language Processing
NPV	Negative Predictive Value
NWIS	NHS Wales Informatics Service
OA	Occupational Asthma
ONS	Office for National Statistics
OPCS	Office of Population Censuses and Surveys

PAD	Peripheral Artery Disease
PEDW	Patient Episode Database
PheKB	Phenotype Knowledge Base
QOF	Quality and Outcomes Framework
RCGP	Royal College of General Practitioners
RCN	Royal College of Nursing
RCP	Royal College of Physicians
RESC	Research Ethics Sub-Committee
SAIL	Secured Anonymized Information Linkage
SCAD	Stable Coronary Artery Disease
SLE	Systemic Lupus Erythematosus
SMR	Scottish Morbidity Records
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SSc	Systemic Sclerosis
SUS	System Usability Scale
SVM	Support Vector Machine
T2DM	Type 2 Diabetes Mellitus
THIN	The Health Improvement Network
TNM	Tumour Node Metastases
UCL	University College London
UKTC	United Kingdom Terminology Centre
UMLS	Unified Medical Language System
WHO	World Health Organisation

Chapter 1

Introduction

The purpose of this chapter is to introduce the topic and context of this thesis, the study setting and population, and the study statement of the problem. It also describes the study's relevance and possible contribution to the broader field of health informatics research, followed by the study's aims and objectives, and the theoretical and methodological approaches that were used in the study.

1 CHAPTER 1: INTRODUCTION

1.1 THESIS TOPIC

Concept libraries are web-based data portals for storing, managing, sharing, and documenting clinical code lists, which are aggregated from linked electronic routine data by using phenotyping algorithms. This thesis informs the requirements for concept libraries for users from a variety of disciplines, including researchers, clinicians, machine learning experts, and research managers, in order to improve disease phenotyping when retrieving and analysing the codes that are collected from linked electronic routine data. Exploratory sequential mixed methods were used in this thesis to look at which concept libraries are available, how they are used, what their characteristics are, where there are gaps, and what needs to be done in the future from the point of view of the people who use them.

1.2 CONTEXT OF THE THESIS

As my home country, the State of Kuwait, was in the process of integrating electronic data (the databases of primary health care centres with hospitals), I travelled to the United Kingdom to pursue a PhD that would be valuable to my country in this context. Therefore, I chose to study more about data linkage and the health informatics that surrounds it. I chose to look at something that was very new in Wales, which was a concept library, and I wanted to examine if this would be valuable to develop in Kuwait. As a result, I began to examine the current concept libraries to learn about their design, contents, and applications in order to assess whether or not users valued them and what a concept library should look like in order to be useful.

1.3 STUDY SETTING AND POPULATION

This research was conducted in Wales, the United Kingdom. This work was an evaluation from the user's perspective of the value of existing and newly proposed concept libraries. First, I invited by email six participants from a variety of disciplines, including researchers, a clinician, a machine learning expert, and a senior research manager in health data science at Swansea University and Cardiff University, to participate in one-to-one interviews, which took place either in their offices or a convenient and private location in the Data Science Building, Medical School, Swansea University campus. Second, I sent an email to all researchers in Wales (N = 34) who work with the SAIL databank, a national e-health data linkage infrastructure, inviting

them to participate in a focus group at the Data Science Building, Medical School, Swansea University campus. Third, I sent e-mail invites to participants working with data linkage centres such as CPRD (Clinical Practice Research Data Link) (N= 200), including researchers, health professionals, and clinicians, requesting them to complete a brief e-mail survey.

1.4 STATEMENT OF THE PROBLEM

The availability of clinical codes in EHR-based research is essential for reproducible research and effective study comparison. The issue is that not all researchers share the code lists they used (e.g., how they were established and the accurate phenotype definitions along with the original research that used them). Although some researchers publish their phenotyping algorithms, there are many challenges associated with effectively reusing and replicating them. For example, the interpretation or manipulation of data often requires knowledge of complex programming languages, such as SQL. This means that EHRs are still inaccessible to many researchers.

To address these issues, the United Kingdom and other countries, including Canada, have built web-based data portals for phenotypes known as concept libraries, which allow data analysts, researchers, and clinicians to upload and download lists of clinical codes, update previous code lists, and share clinical code data across platforms. Also, a team of health informatics and data scientists in Swansea University and Cardiff University with clinicians and statisticians has built a prototype concept library in the SAIL databank for diagnoses, symptoms, and medications.

Despite the fact that the development of concept libraries appears to be excellent for addressing all of the mentioned problems, existing concept libraries are not widely used. The purpose of this thesis was to examine the barriers that prevent researchers from sharing their clinical code lists or reusing others' code lists in existing concept libraries and to explore the needs of various users, including researchers, clinicians, and managers, in the development of Swansea University's prototype concept library.

1.5 SIGNIFICANCE OF THE STUDY TO THE FIELD OF HEALTH INFORMATICS

To our knowledge, this is the first PhD thesis aimed at identifying the needs of various users of a concept library. The findings of this study would have a significant impact on improving the efficiency of existing concept libraries by informing their developers about the different

requirements, facilitators, barriers, and recommendations of the various users. This work will greatly inform the developers of new concept libraries to improve access to and collaboration with EHRs' routine data, which is part of an all-UK agenda, and the findings of this study will have implications for other countries working to access and share linked EHRs' routine data, such as the State of Kuwait.

In addition, this thesis includes two reviews of the literature: the first review aims to explore existing concept libraries, examine their utilities, identify the current gaps, and suggest future developments; the second review aims to summarise the clinical classification systems used to identify chronic diseases in children in routine data sources and other administrative datasets**,** which serves to illustrate an example of how a concept library could be developed to give 'modules' or concepts in specific clinical areas. In addition, This review may also be useful for researchers interested in examining studies that classify a range of chronic conditions in children using a range of different combined data sources and beneficial for existing concept libraries as I tried to make the classification systems presented in this review repeatable by summarising the definitions, types of routine data sources, coding systems on which they are based (for example, ICD-10), and links to their specific codes if they are publicly available.

1.6 AIMS AND OBJECTIVES

This thesis aims to explore the requirements of various users, including researchers, clinicians, machine learning experts, and managers for the development of a data portal for storing, managing, sharing, and documenting clinical code lists, referred to as a concept library, and to examine why existing concept libraries are not widely used.

1.6.1 The thesis specific aims

1) To investigate and summarise the various features of existing concept libraries in the literature, such as definitions, types, similarities, and differences. (Chapter 4)
2) To discover the different clinical classification systems used in the literature to identify chronic conditions in children from routine data sources. (Chapter 5)
3) To explore the requirements of various users, such as researchers, clinicians, machine learning experts, and research managers, when using concept libraries (Chapter 6).
4) To examine why current concept libraries are not commonly used. (Chapter 6)
5) To identify the current system for disease phenotyping used by the users (its strengths and weaknesses). (Chapter 6)

6) To develop an email survey to explore the usability of concept libraries by various users (Chapters 3 and 7)

1.7 Research methodology and methods

In this thesis, I used an exploratory sequential mixed methods design that consists of three phases: 1) two qualitative studies, 2) the development of an email survey, and 3) one quantitative study. In the first phase, I conducted the following two qualitative studies:

The first qualitative study involved one-to-one interviews with researchers, clinicians, and managers, which were conducted to examine their specific needs for concept libraries. Based on the purpose of this study, I developed semi-structured interview questions, which consist of introductory, flow, key, and final questions.

The second qualitative study involved a focus group with researchers working with the SAIL databank. It was held for 2 hours, and a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system and the proposed concept library was performed. Before the focus group session, I created a list of ten semi-structured questions based on the objectives of this study. The purpose of the questions was to generate thoughtful and thorough responses from the participants; therefore, I avoided using closed-ended questions (e.g., yes or no).

In the second phase, I developed an e-mail survey instrument based on the findings of the first qualitative phase of this research and called it the Concept Library Usability Scale. This development links the study's first qualitative phase to the subsequent quantitative phase, which represents the point where the two phases mix.

In the third phase, I examined the content consistency and validity of the Concept Library Usability Scale, an e-mail survey instrument that I developed in the second phase. The e-mail survey was tested with several participants to ensure that the questions were understandable to the users of concept libraries, who are the target participants for the survey. I sent e-mail invitations to the participants, which included researchers, health professionals, and clinicians, asking them to complete the Concept Library Usability Scale.

1.8 Structure of the thesis

This thesis informs the requirements for a concept library for users from a variety of disciplines, including researchers, clinicians, machine learning experts, and research managers, in order to

improve disease phenotyping when retrieving and analysing the codes that are collected from linked electronic routine data within the SAIL in Wales.

The purpose of chapter 1 is to introduce the topic and context of this thesis, including the study setting and population, as well as the statement of the problem. Next, chapter 1 describes the study's relevance and possible contribution to the broader field of health informatics research, followed by the study's aims and objectives, and the theoretical and methodological approaches that were used in the study. This chapter concludes by providing an overview of the thesis's structure.

Chapter 2 provides a background to the topic of the thesis. It describes the importance of linked electronic patient data as a resource for conducting research, provides an overview of some of the existing linked electronic patient data repositories in the UK, mentions some of the challenges associated with the identification of EHRs based-phenotypes, and summarises approaches as well as uses of EHRs based-phenotyping.

Chapter 3 details all of the exploratory sequential mixed approaches used in the three phases of this thesis: 1) two qualitative studies, 2) the development of an email survey, and 3) one quantitative study.
The purpose of Chapter 4 is to review the literature on existing concept libraries for disease phenotyping, which serve as platforms for multiple researchers to store, manage, and share phenotypes (diagnoses, symptoms, medications, and procedures). This review aims to examine how they are used and identify current gaps and future development.

Chapter 5 involves conducting a review of the literature to identify some of the existing classification systems used for identifying children with chronic conditions in routine data sources.

Chapter 6 explores the needs for a portal of disease phenotyping definitions (a concept library) for users from a variety of disciplines, including researchers, clinicians, machine learning experts, and research managers. Additionally, it analyses why existing concept libraries are not more widely used.

Chapter 7 presents the results and statistics of the quantitative study that was performed using the Concept Library Usability Scale, an e-mail survey instrument. The purpose of this study

was to investigate the various requirements for a concept library for users from various disciplines, such as academics, clinicians, machine learning experts, and research managers.

Chapter 8 discusses in detail the findings of the three phases of this thesis: 1) the two qualitative studies, 2) the development of an email survey, and 3) the quantitative study. It also addresses the thesis's strengths, limitations, and original contributions to concept libraries. This chapter also describes the identified opportunities and obstacles from the users' perspectives in order to maximise the utility of concept libraries. Finally, it gives recommendations for future research directions in order to improve reproducible research.

Chapter 9 presents the conclusion of the overall studies, including the interviews, focus groups, and survey conducted in this thesis, and provides recommendations to improve repeatable research using linked routine electronic data sources.

Chapter 2

Background

This chapter provides a background to the topic of the thesis. It describes the importance of linked electronic patient data as a resource for conducting research, provides an overview of some of the existing linked electronic patient data repositories in the UK, mentions some of the challenges associated with the identification of EHRs based-phenotypes, and summarises approaches as well as uses of EHRs based-phenotyping.

2 CHAPTER 2: BACKGROUND

Health care systems are becoming more digitally focused rather than paper-based and are moving to the use of electronic health record (EHRs). The growing availability of electronic patient data offers health care practitioners increased opportunities for secondary use of EHRs data to improve the quality of care and research [1] [2] [3]. This chapter describes the importance of linked electronic patient data as a resource for conducting research, provides an overview of some of the existing linked electronic patient data repositories in the UK, mentions some of the challenges associated with the identification of EHRs based-phenotypes, and summarises approaches as well as uses of EHRs based-phenotyping.

2.1 LINKED ELECTRONIC PATIENT DATA

The availability of large amounts of electronic patient data that can be moved and linked together into safe data repositories could enable researchers and data analysts to query and examine this data effectively [4] [5] [6]. There has been an annual rise at a rate of approximately 20 % in primary care research using EHRs in the United Kingdom (UK) [7], which gathers data about general practice from the following databases: 1) Clinical Practice Research Data Link (CPRD) [8] 2) The Health Improvement Network (THIN) [9] 3) QResearch [10], and 4) Secured Anonymized Information Linkage (SAIL) [11].

About 98% of the UK population are enrolled with a primary care general practice (GP) [12] and under the National Health Service (NHS), GP visits are available for free. The GP is the care gatekeeper of the UK NHS. GPs serve as the first point of contact with any non-emergency health-related problems that can then be handled within primary care and/or referred to secondary care as appropriate. Secondary care teams also provide GPs with knowledge about their patients, including main diagnoses [8].In addition, secondary care data such as the hospital admission system (HES – England, PEDW – Wales, SMR -Scotland), are linked to primary care records [6] [7] [13] [14]. Such linked information creates the opportunity to undertake research into the causes and outcomes and pathway of disease.

2.2 EXAMPLES OF LINKED ELECTRONIC PATIENT DATA REPOSITORIES

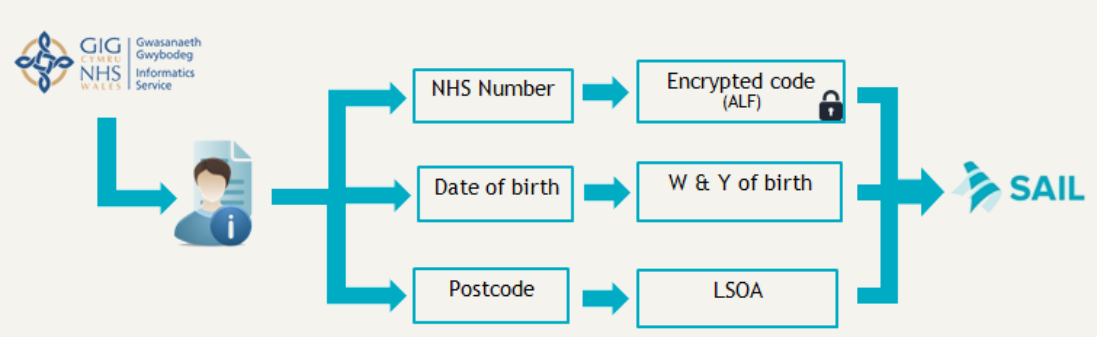
2.2.1 The Secure Anonymised Information Linkage (SAIL)

SAIL is a national data safe haven with datasets, primarily on the population of Wales. It contains information on over 5 million people who have received public services in Wales, which are made accessible anonymously to researchers around the globe. SAIL was developed by Swansea University (Wales, UK) Population Data Science Group in 2007 with core funding from the Welsh Government's Health and Care Research Wales (HCRW). It was designed to optimise opportunities for longitudinal, cross-sectoral assessment of services and interventions while at the same time protecting privacy [13].

SAIL has a collection of Wales-wide, core or core-restrictive datasets. The Independent Information Governance Review Panel (IGRP) can authorise the use of both datasets. The difference is that only data providers reserve the right to review the proposed uses of core-restricted datasets. Additionally, SAIL can integrate study datasets generated by researchers, providing they have obtained all necessary regulatory approvals, including consent to permit them to transfer the dataset to SAIL and connect it to existing SAIL data. Access to a study data set may be restricted, depending on regulatory approvals, to a specific project team or individual researcher [13].

SAIL datasets are linked through the use of an Anonymous Link Field (ALF): a special, anonymous identifier assigned to every person in a dataset by NHS Wales Informatics Service (NWIS) after a matching process against the Welsh Demographic Service Database (WDS) database. Figure 2.1 shows the process of data anonymisation. In this figure, W& Y stand for Week and Year of Birth and LSOA stands for Lower Layer Super Output Areas. NWIS and SAIL co-designed and tested the matching algorithm. First, SAIL focused on determining the appropriateness of using the NHS number as the basis for a unique identifier, and then to evaluate the algorithm to match a collection of datasets from primary care, secondary care and social services against WDS. Also another benefit of the matching process in NWIS is that ALFs can be assigned to datasets originating from outside the health sector (such as social services or education) or where the NWIS number is otherwise lacking [11] [13].

Figure 2.1: NHS Wales informatics service and data anonymisation



Source: https://saildatabank.com/faq/

SAIL has more than 1,200 registered data users, and more than 300 projects have used the data, with about 30 new projects each year. For example, research using SAIL data has been used to inform National Institute for Health and Care Excellence (NICE) guidelines and European Guidelines for the treatment of patients with ankylosing spondylitis (AS). Previously patients could only be given disease-modifying medications if they had serious disease for at least 3 months. However, SAIL research showed that flares last about a month, and that there is no need to postpone treatment for people beyond this time. The study measured the cost of treating AS in the UK to inform service planning [15] [16].

2.2.2 The CALIBER research platform – electronic health records for cardiovascular research

The CALIBER research platform is "*a unique research platform consisting of 'research ready' variables extracted from linked electronic health records (EHR) from primary care, coded hospital records, social deprivation information and cause-specific mortality data in England"* [17].

It is led by University College London (UCL) and partners including UCL, the London School of Hygiene and Tropical Medicine, and Queen Mary University of London. CALIBER investigators are a group of epidemiologists, clinicians, statisticians, health informaticians, and computer scientists who work together [18].

Data in CALIBER platform (https://www.caliberresearch.org) are "research ready" variables derived from four linked NHS electronic health records and administrative health data that are deterministically connected using the NHS number (a specific 10-digit identifier assigned at birth or first interaction), gender, postcode, and date of birth [19] including: the Clinical Practice Research Datalink (CPRD) [20], the Myocardial Ischaemia National Audit Project (MINAP) [21], Hospital Episodes Statistics (HES) [22], and the Office for National Statistics

(ONS) [23]. Types of data, coding system used and data recording details of these linked electronic health record sources presented in Table 2.1[18].

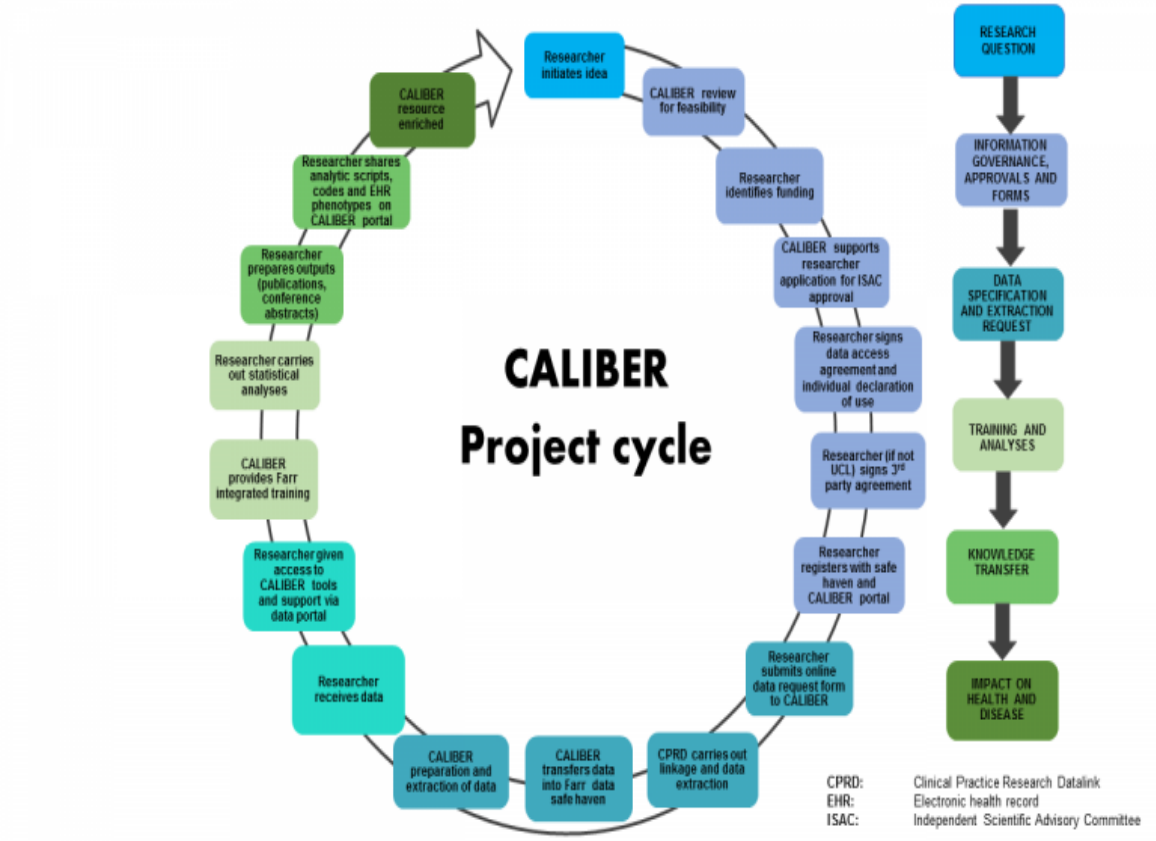
Table 2.1: Linked electronic health record sources in CALIBER

Sources	Types of data	Coding system	When and by whom data is coded?
Primary care: CPRD and other sources	Longitudinal primary care data Diagnoses and symptoms irrespective of hospitalization, drug prescriptions, vaccinations, blood test results, risk factors	Data recorded using the Read clinical terminology system, version 3 contains ~99 000 codes	Data recoded by the general practitioner in real time during the consultation Hospital discharge letters coded by a practice administrator
Social deprivation: ONS	Small area patient social deprivation data	Index of Multiple Deprivation (2007) and Townsend score	Derived from multiple national administrative data sets
Disease registry: MINAP	National registry of Acute Coronary Syndrome admissions Phenotype (ST Elevation Myocardial Infarction, Non-ST Elevation Myocardial Infarction, Unstable Angina), severity and treatment data	In all, 120 fields most with multiple response categories, as defined by the MINAP steering group	Recorded usually by audit nurse, days or weeks after admission, by abstracting data from hospital records
Secondary care: HES	National data warehouse of hospitalizations recorded for administrative purposes Inpatient, outpatient, emergency, critical care and maternity admissions[a] Operations and surgical procedures	Up to 20 primary and secondary discharge diagnoses recorded using ICD-10 Up to 24 codes using the Office of Population, Censuses and Surveys Classification of Surgical Operations and Procedures and used for operations The 4th revision (OPCS-4) contains ~10 000 codes	Recorded by non-clinical trained coders based on the discharge summary weeks after discharge
Mortality: ONS	National census of all deaths Primary and underlying cause of death	The primary, underlying and up to 14 secondary causes of death are recorded using ICD-10	Doctor (general practitioner or hospital) completes death certificate with cause of death. ICD codes added by trained non-clinical coders

Source: (Denaxas, 2012, P.1627)

Researchers can use raw data after the protocol has been approved by the bodies that govern access to the constituent data sets and payment has been made to them. For CPRD, this includes the Independent Scientific Advisory Committee's (ISAC) scientific approval of the protocol, as well as a signed license specifying the nature and data confidentiality of CPRD data use. For MINAP, applications are made to the MINAP Academics Group, and for HES and ONS, applications are made to the NHS Digital, which is the national information and technology partner to the health and care system [24]. An application should be submitted to the Scientific Advisory Committee of CALIBER following such approvals [18]. Steps involved in accessing CALIBER resources summarised in the CALIBER project cycle (Figure 2.2) [17].

Figure 2.2: CALIBER project cycle



Source: (UCL Institute of Health Informatics)

CALIBER resources contribute new information across different stages of translational pathways such as: 1) Risk factors for the onset of cardiovascular diseases (CVDs): CALIBER cohorts allow research into the early stages of a wide range of CVDs, which was previously impossible in smaller, less clinically phenotype cohorts. Smoking, for example, has varying associations with different CVDs endpoints, with no correlation with initial presentation with ventricular arrhythmia or cardiac arrest, a modest association with chronic stable angina and heart failure, and significant associations with Acute Coronary Syndromes (ACSs) [25]. 2) Quality of treatment and outcome studies: There are Myocardial Infarction (MI) registries in two countries around the world, which provide data on consecutive patients in all hospitals. There were significant care disparities in comparisons of over 500,000 patients in Sweden and the UK, and 30-day mortality was lower in Sweden [26]. 3) Prognosis: Prognostic models have been developed and validated in a linked electronic health records cohort of 102 023 patients

with Stable Coronary Artery Disease (SCAD) for estimating risk of all-cause mortality and coronary outcomes based on clinical criteria that are typically available in all people with Stable Coronary Disease (CAD) [27].

2.2.3 The Clinical Practice Research Datalink (CPRD)

The CPRD is an active database of general practitioners' anonymized medical records for primary care [8]. Anonymized primary care data for patients can be linked individually to secondary care and other health and area-based datasets. This linkage allows CPRD to have a better understanding of the patient care record to facilitate critical public health studies, inform improvements in patient safety and care delivery [28]. 10.6 million patients out of 411 in general practise in the UK participated in the record linkage in June 2018. In the CPRD standard linked dataset release, 9.1 million (86 percent) patients were of research quality, of which 8.0 million (88 percent) had a valid NHS number and were eligible for linkage [20].

CPRD has set up a data linking programme that routinely connects primary care data to other patient-level health data from data custodians NHS Digital and Public Health England (PHE). Data linkages are made NHS Digital, acknowledged in law as by the Health and Social Care Information Centre (HSCIC), the national provider of health and social care information, data and IT systems in health and social care in the UK, and the legislative body in the UK legally allowed to collect identifiable patient data. CPRD subsequently collects and delivers anonymous linked data and metadata for the researchers including data from Hospital Episode Statistics, Office for National Statistics, and National Cancer Registration and Analysis Service [20]. CPRD routine linkages are shown in Table 2.2.

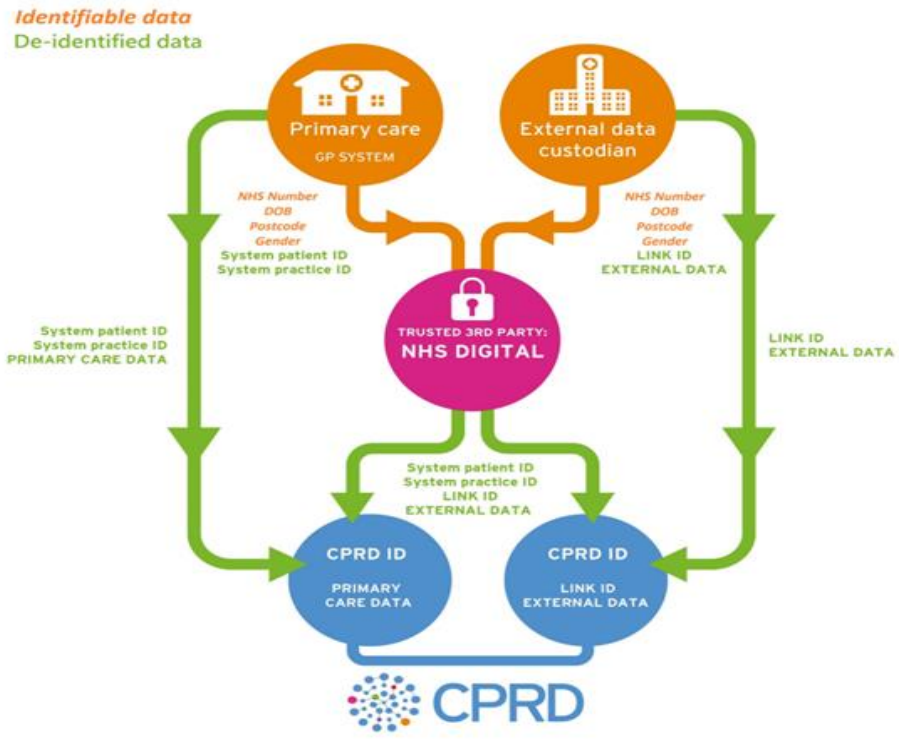
Primary care data are provided to CPRD by general practice electronic software providers functioning as data processors. Data are provided on a regular basis from practices that have consented to share data (Figure 2.3). Personal identifiers including name, full date of birth, postcode and National Health Service (NHS) number are deleted at source by the system provider and substituted by pseudonymized system patient and practice identifiers before to the transmission of data to CPRD [20] .

Table 2.2: CPRD routine linkages

Hospital Episode Statistics Admitted Patient Care (HES APC)
Hospital Episode Statistics Outpatient (HES OP)
Hospital Episode Statistics Accident and Emergency (HES A&E)
Hospital Episode Statistics Diagnostic Imaging Dataset (HES DID)
Office of National Statistics (ONS) Death Registration
National Cancer Registration and Analysis Service (NCRAS) data from Public Health England (PHE) including:
Cancer registration data
Cancer Patient Experience Survey (CPES) data
Systemic Anti-Cancer Treatment (SACT) data
National Radiotherapy Dataset (RTDS)
Mental Health Dataset (MHDS) data
Measures of relative deprivation and rural urban classification at Lower Layer Super Output Area (LSOA) level for practices and patients

Source: (Padmanabhan, 2019, P.92)

Figure 2.3: Flow of primary care data and external data custodian



Source: (Padmanabhan, 2019, P.92)

External data custodians send to NHS Digital their patients' personal identifiers including NHS number, gender, date of birth and postcode, and a pseudonymised patient record identifier (Table 2.3). NHS Digital matches identifiers provided by external data custodians to primary care identifiers in the CPRD cohort file, creating a linker file. The linker file includes a pair of pseudonymized identifiers (GP system patient and practice ID, external dataset link ID) for each connected patient which can be used to integrate the primary care dataset with the external dataset [20] .

Table 2.3: Deterministic linkage steps

Step (match rank)	Match required
1	Exact NHS number, gender, DOB and postcode
2	Exact NHS number, gender and DOB
3	Exact NHS number, gender, postcode and partial DOB
4	Exact NHS number, gender and partial DOB
5	Exact NHS number and postcode
6	Exact gender, DOB and postcode
	(NHS number must not contradict the match, DOB must not be 1st of January and postcode must not be on the communal establishment list[a])
7	Exact gender, DOB and postcode
	(NHS number must not contradict the match and DOB must not be 1st of January)
8	Exact NHS number

[a]Communal establishments include: hospitals, care homes, prisons, defence bases, boarding schools and student halls of residence

Source: (Padmanabhan, 2019, P.92)

Currently, more than two-thirds of the protocols for access to CPRD data require primary care data combined with other health-related data sets in order to extend research scope and increase validation of study results. For example, Herrett et al. showed a substantial improvement in the identification of myocardial infarction using linked primary care, HES, Office for National Statistics (ONS) and Myocardial Ischaemia National Audit Project (MINAP) data, with single sources underestimating rates of up to 50% [29].

2.3 CLINICAL CODING

Clinical coding systems are standardised, and often hierarchical, lists of medical terminology that can be used to identify patient procedures, symptoms, diseases, results and treatments [30]. A clinical coding can be defined as *"the translation of medical terminology that describes a patient's complaint, problem, diagnosis, treatment or other reason for seeking medical attention into codes that can then be easily tabulated, aggregated and sorted for statistical analysis in an efficient and meaningful manner"* [31]. A clinical coder is a health informatics professional who translates the medical terminology into classification codes in a medical record of a patient. Clinical coders use their expertise, knowledge and experience to allocate codes precisely and reliably in compliance with the classification and national clinical coding criteria [31].

Although there is a belief that the primary objective of clinical coding is for reimbursement claims [32], there are several advantages to the deployment of the coding systems. Clinical coding systems facilitate precise recording of the conditions of the patient, which can lead to improved medical decision-making even in situations where the condition is complicated. In addition, structured coding systems facilitate the clinical classification of morbidity and mortality for statistical purposes and the determination of trends of study variables. Clinical coding systems are methods for documenting a clinical incident in a manner that can be easily accessed, obtained, arranged, sorted, processed, filtered and transmitted [30].Two distinct sets of coding systems used in healthcare are clinical terminologies and clinical classification systems [33].

2.3.1 Clinical terminologies

A clinical terminology is a standardised set of descriptive terms used at the point of care in clinical practices. It can include diagnoses, procedures, medications, and administrative data [34]. Imel and Campbell defined a clinical terminology as *"a set of concepts and relationships that provide a common reference point for comparisons and aggregation of data about the entire health care process, recorded by multiple different individuals, systems, or institutions"* [35].

Clinical terminology systems use a conceptual approach to the arrangement of labels within a domain and are also used as a standard reference framework for mapping between systems. These systems perform well in predictive analysis and international studies. Concepts may be

merged in terminology systems to create complicated nomenclatures or naming systems [30].The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Read codes are examples of clinical terminology systems used in healthcare.

2.3.1.1 Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)

SNOMED CT was developed by the International Health Terminology Standards Development Organization (IHTSDO) [30]. It offers a terminology for documenting structured data in electronic records relating to the health and treatment of the patient. It allows the user to record and understand standard meaningful clinical terms and facilitates advanced computer interpretation, and it offers features that allow the functionality of decision support and powerful analytics. The use of SNOMED CT would also allow sharing of data across various systems in the health-care field [36].

SNOMED CT is used in many different healthcare facilities all over the world, including enhancing breast cancer treatment in Sweden and controlling allergies in Buenos Aires. In UK, all GP systems use SNOMED CT to capture clinical terminology. It is used across health facilities in different systems and services such as the Quality and Outcomes Framework (QOF), identifying patients for flu vaccinations, Electronic Referral Service (e-RS), and Summary Care Records [37].

The use of SNOMED CT as a standard language for documenting clinical patient information across the NHS ensures that patient information is recorded regularly and precisely, which facilitates the sharing of clinical information among systems. For instance, clinical data in the discharge summary may be directly integrated into the GP patient record without the need for the care provider to re-enter the data manually, not just saving time, but also eliminating the risk of human error. Northern Ireland, Scotland and Wales have work projects ongoing related to SNOMED CT to capture clinical terminology in electronic patient record systems [37].

SNOMED CT is provided as a series of data files. For the application of the standard for healthcare terminology, these data files must be integrated into the electronic system. The application of the standard must be done as part of a digitalized initiative at the point of care. The standard is required for a wide range of applications such as electronic patient record system, electronic health records, electronic care plans, Specialist systems such as Cancer

MDT systems, Decision support tools, Clinical Knowledge Resources, Clinical Guidance, Care Pathways, and Messages between care solutions. The data items to be registered using SNOMED CT include but not limited to diagnosis, procedures, assessment scales, family history, medications, allergies, observations such as blood pressure, and documentation type and documentation care setting [36].

2.3.1.2 Read codes

Read terms are a collection of clinical definitions that can be used to manage data in patient records. They are named after Dr. James Read, the original author of these words. Read Terms relates to descriptions such as Asthma. Each term is also paired with the Read Code, which is the letter and number code that uniquely defines the clinical term. It can be up to five characters in length, including full stops. For example, the 'H33..' Read Code is associated with the 'Asthma' Read Term [38].

Read codes are used in the NHS since 1985. Two versions are available: version 2 (v2) and version 3. (CTV3 or v3). These versions provide clinicians with a standard vocabulary for recording patient outcomes and procedures, in health and social care IT systems in primary and secondary care [39].

Read codes have known problems that cannot be solved, for example, they have incorrect content, no space for new content at the correct place and ideally require over 5 characters to represent such content correctly. Read codes cannot even be extended to meet the needs of all healthcare providers, and they are only available in the United Kingdom, creating a challenge for international cooperation in fields such as rare diseases and genomics. Healthcare models rely more and more on data sharing across domains and therefore a standard terminology is essential for their potential success. Advanced terminology is needed for EHR systems to provide useful support to future healthcare, such as an effective analysis of the data. In the UK, consultation with the Joint GP IT Committee (the BMA and RCGP) and the various clinical bodies (e.g., AoMRC, AHPs, RCN, RCP) led to the agreement that the transformation from Read code to SNOMED CT is essential. All suppliers of systems are required to assign appropriate SNOMED CT code for each Read term already in medical records using national mapping tables provided by the United Kingdom Terminology Centre (UKTC). In this context, a "map" means a connection between the term in Read and the related definition in SNOMED CT. This means that there is a consistent approach across the various GP systems to extract the resulting SNOMED CT code for each Read code [40].

2.3.2 Clinical classification systems

Clinical classification system is the arrangement of clinical terms in categories [41]. Clinical classification systems are structured to classify clinical conditions and procedures in order to facilitate statistical data analysis across the healthcare system, and to provide criteria for the comparison of national and international health statistics [33]. The UK NHS has a long history of using classification systems such as ICD-10 and OPCS-4 in order to track the health of UK people and conduct business processes including payment. They are used also to categorise the completed episode of treatment according to predetermined classification codes and to endorse indirect care related practices such as epidemiology, payment and population surveillance [36].

2.3.2.1 The International Classification of Disease (ICD)

The ICD is a comprehensive index of diseases and injuries that is developed and published by the World Health Organisation (WHO). ICD is the basis for health statistics and the international standard for monitoring disease patterns and health conditions. It is the standard classification used in clinical and study purposes. The ICD classifies diseases, disorders, injuries and other associated health problems, listed according to a systematic, hierarchical structure that enables: 1) simple storage, retrieval, and review of health data for evidence-based decision making, 2) distribution and comparison of health data among hospitals, regions, facilities and countries, and 3) comparison of data at the same location over different time periods. Based on clinical information, studies, and epidemiological data, ICD has become a useful health tool that is appropriate for many health uses such as: 1) control of the incidence and prevalence of diseases, 2) causes of mortality, 3) external causes of the disease, 4) primary care facilities, and 5) registration of rare diseases [42].

The first international classification version (the International List of Death Causes), was established by the World Statistical Institute in 1893. The ICD has been updated and released in several revisions, reflecting developments in health and medical science. The WHO was trusted with the ICD at its beginning in 1948 and published the sixth version, ICD-6, which included morbidity. The WHO Nomenclature Regulations, introduced in 1967, mandate Member States to use the most up-to-date revision of the ICD for national and international registration and reporting data on mortality and morbidity statistics. ICD-10 was endorsed in May 1990 by the Forty-third World Health Assembly. It is cited in more than 20,000 scientific articles and used by more than 150 countries around the world and has been translated into more than 40 languages. With the need for more detailed recording and reporting, a number of

clinical modifications or specialty adaptations proliferated over time. ICD-11 was approved by the Seventy-second World Health Assembly in May 2019. ICD-11 reunites the various modifications and adaptations, adds clinical needs, and converts ICD from a statistical framework to a clinical classification for statistical use. [42].

2.3.2.2 Office of Population Censuses and Surveys (OPCS) for procedures and interventions

The Office of Population Censuses and Surveys for procedures and interventions is a statistical classification of interventions and procedures carried out in the NHS [43] . The fourth revision of the OPCS (OPCS-4) was initially published in 1987, with final release and implementation in 1990. The general goals of the review process were:

1. To define and describe existing surgical operations with special regard to the implementation of recent advanced techniques.

2. To exclude seldom conducted operations but to include procedures that do not require a complete operating theatre environment.

3. To provide flexible classification, adaptive to less identified specialty parameters and capable of potential expansion [31].

OPCS-4 is a mandatory requirement for the NHS Admitted Patient Care Commissioning Data Sets in the United Kingdom. OPCS-4 supports different types of data collection such as Central Returns and Commissioning Data Sets (CDS). All consultant episodes including procedures must be registered and collected using OPCS-4. Requirements for data sets and relevant descriptions are defined in the NHS Data Model and Data Dictionary. The OPCS-4 also supports different types of secondary use of information that are important for the planning and improvement of patient care. These secondary uses include operational and strategic planning, resource utilisation, national and local planning and performance management, research and epidemiology, and the NHS payment system [31] [43].

Health care providers who have adopted electronic health records and clinical terminology such as SNOMED CT use the connection between the terminology and OPCS-4 known as 'cross-maps' to allow the clinical coding of electronic health records. The national cross-maps comply with national clinical coding guidelines. They are available in the biannual updates of the NHS Digital UK SNOMED CT Clinical Edition. They are intended to encourage all organisations with electronic health systems to comply with the required obligation for the collection and

reporting of intervention and procedure data using the NHS Information Standard, OPCS-4 [31].

2.3.2.3 National Interim Clinical Imaging Procedure (NICIP) Code Set

The NICIP code set is *"a comprehensive national standard set of codes and descriptions for imaging procedures and is maintained by the UK Terminology Centre"* [44]. The NICIP code set includes a consistent definition of imaging procedures and a standard terminology for the description of aspects of clinical imaging procedures to identify the procedures performed in an image review and to provide clinical details on the identified procedures [44]. The NICIP code set is developed to fill the gap before SNOMED CT is natively supported by all clinical systems. When all clinical systems are using SNOMED CT, it is expected that clinical imaging procedures will be represented entirely by the use of SNOMED CT coding concepts in NHS Care Record Service (NCRS) applications [45].

The NICIP code set was created as a collaborative project with structured management agreements and with the engagement of all known key stakeholders. A group has been established to manage developments – the Clinical Imaging Management Group (CIMG). This management group is actually the editorial scaling-up council, the advisory group and the regulatory body for NICIP codes and related products, while NHS Digital takes the maintenance and update responsibilities. The CIMG recommends that the NICIP code set be used in all integrated clinical application systems. This strategy is supported by all stakeholders, including the Royal College of Radiologists, British Nuclear Medicine Society, British Medical Association, Royal College of General Practitioners, the Society of Radiographers, the Department of Health and Social Care, NHS Digital, NHS England, and Delegates of England, Scotland, Wales, and Northern Ireland [45].

2.4 ELECTRONIC HEALTH RECORDS-BASED PHENOTYPING

Identifying patients with particular conditions or results (i.e., patient cohort identification from the clinical codes), known as phenotyping, is the basis for translational research, comparative efficacy studies, clinical decision support, and population health analyses using regularly collected EHRs data [46]. The different approaches to EHRs based-phenotyping are the following:

2.4.1 Rule-based approach

A rule-based approach or machine-based approach is typically used to classify a group of patients with particular selected characteristics. In a rule-based technique, a collection of rules on extracted features must be established to classify the records of patients. Machine-based technique, however, use computer programs that can learn by themselves, identify data patterns and adjust actions according to new data [47].

A rule-based approach is a collection of conceptual criteria (rules) which could be applied for EHRs data to assess patients' phenotype status, for example, haemoglobin <10 AND age>60) [46] [51]. A rule-based approach is simple to use and easy to create, precise in the use of small datasets, and reliable in that it uses human-interpretable techniques. The performance of rule-based approaches is better when phenotypes have simple procedural and diagnostic codes [46]. However, because of the need for clinical and informatics expertise, the application of this approach requires time, commitment and effort [47].

Rule-based algorithms for phenotyping vary from basic pattern matching to more complex symbolic approaches. They can involve several logical steps and may incorporate a variety of operations such as Boolean (AND, OR and NOT), Comparative (threshold variable) and Aggregative (COUNT, FIRST) [48]. There are two different techniques to generate the rules for building rule-based phenotyping: 1) rule based on clinical judgment and 2) rules based on healthcare guidelines.

1) Rule based on clinical judgment

In this technique, one or more clinicians are involved in defining inclusion and exclusion criteria (logical rules) based on standardised data elements such as diagnostic codes, medications, procedures and laboratory values [49]. For example, Michalik et al. have used electronic health record data for the creation of a computerised algorithm for Sickle cell disease

from the Children's Hospital Wisconsin. The algorithm was based upon the International Statistical Classification of Diseases, the Ninth Review Codes, number of visits and sickle-cell hospital acceptances. The algorithm was improved in an iterative method with the aid of computer technology to combine biology and bedside queries. A manual medical evaluation and a gold standard set of confirmed sickle cell cases were used to verify the final algorithm. Then, the algorithm was validated at the Froedtert Hospital, a neighbouring adult health system [50]. Nguyen et al. developed a symbolic rule-based classification scheme to automatically classify lung tumour node metastases (TNM) cancer from free-text pathology. They used a tool called MEDTEX, which includes modules for mapping free text to terms of Systematized Medicine Clinical Terms Nomenclature (SNOMED-CT), negation finder, and possibility recognition. Their system performance was comparable to support vector machine (SVM)-based text classification systems [51]. Restrepo et al. developed an algorithm using data mining methods to identify a cohort of type-2 diabetic African Americans de-identified electronic medical records (EMRs) from the Vanderbilt University. The algorithm involves a combination of diagnostic codes, existing procedural terminology billing codes, medications, and matching text to classify Diabetic Retinopathy (DR) if gold-standard digital photography results were unavailable. DR cases were reported with 75.3% positive predictive value and 84.8% precision. Controls were rated as a negative predictive value of 1.0 percent [52].

2) Rules based on healthcare guidelines

Some research uses health institutions' guidelines to derive rules for particular diseases. Aref-Eshghi et al. reviewed EMRs of patients visiting primary care clinics in St. John's, Newfoundland and Labrador (NL), Canada in 2009–2010 to determine the best dyslipidaemia recognition algorithm. The authors developed six algorithms containing three components, dyslipidaemia ICD coding, lipid-lowering drug use and abnormal laboratory lipid levels, were measured against a gold standard, which was identified as the existence of any of the three criteria [53]. Kho et al. constructed an algorithm to classify cases and controls of type 2 diabetes using routinely collected multi-institutional EHRs data. An algorithm based on diagnostic codes, medications, and laboratory test results was developed using existing clinical diagnostic criteria provided by the American Diabetes Association. Small improvements were made to ensure the portability of the algorithm across multiple organisations. The authors attribute PPV of their algorithm to several iterations and chart analysis [54]. Safarova et al. have created an EHRs-based phenotyping algorithm to rapidly classify familial hypercholesterolemia (FH) in

the Screening Employees and Residents in the Community for Hypercholesterolemia (SEARCH) study. The Dutch Lipid Clinic Network requirements have been calculated with an EHRs-based algorithm using standardised data sets and natural language processing for the family history and existence of FH stigma during physical examination. A blind expert analysis showed positive and negative SEARCH predictive values of 94% and 97%, respectively [55].

2.4.2 Machine learning approach

Machine learning (ML) based phenotyping algorithms allow programs to derive patterns from a dataset during the learning process, enabling them to generalise predictions across multiple datasets. During the training phase, numerical parameters representing the underlying structure of a particular algorithm are optimised using several iterative methods. The use of ML approaches in phenotyping algorithms will minimise the effort required by clinical experts, as there is no need to create rules. ML approaches are used in this context to enable accurate prediction of target diagnoses on the basis of the observations of relevant samples. ML approaches are categorised as supervised or unsupervised [47].

1) Supervised machine learning

In supervised machine learning, the computer is provided with observed training data and related identified output values. The aim is to learn general rules (also called a "model") that map the inputs to outputs, so that the output for new invisible data can be predicted when input values are detected, but not their related output [56]. The major issue with supervised machine learning approaches is the probability of over-fitting (i.e., The model works for matching the data of the learning samples accurately, but it does not predict the unseen data well). The over-fitting occurs when the model learns random noise during the learning process rather than just from the desired functionality. The possibility of overfitting can be minimised by implementing a cross-validation methodology that helps provide more reliable performance assessment in unseen cases [47]. There are several types of supervised machine learning approaches that were used in the literature including: 1) support vector machine (SVM) 2) decision tree, and 3) logistic regression approaches.

1) Support vector machine (SVM)

The SVM is a machine learning algorithm that is used to learn classification and regression models. Given a dataset with two linearly separable target classes, it turns out that there are an infinite number of lines that can distinguish between the two classes. The SVM's objective is to determine the best line that divides the two classes. This line is known as a hyperplane in higher dimensions [57]. The SVM based approach was used in previous studies. For instance, Zeng et al. developed a new concept-based filter and a prediction model to detect local recurrence of breast cancer using EHRs. The authors manually analysed a development corpus of 50 progress notes in their training dataset and extract partial sentences that show local recurrence of breast cancer, and they used MetaMap to process these partial sentences to get a collection of Unified Medical Language System (UMLS) concepts. These features are used together with the number of pathology reports reported for each patient to train SVM for recognising local recurrence. They compared their model to three basic classifiers using either complete MetaMap concepts, filtered MetaMap concepts, or word bags. Their model has achieved the highest Area under receiver operating characteristic curve (AUC) (0.93 in cross-validation, 0.87 in held-out testing).[58].

2) Decision tree

A decision tree is a tree in which the internal nodes can be interpreted as tests (based on input data patterns) and the leaf nodes can be interpreted as categories (of these patterns). The tests are filtered through the tree in order to obtain the correct output to the input pattern [59].Several decision tree algorithms have been used in the literature, for example, Zhou et al. used a machine learning scheme to classify patients with rheumatoid arthritis from primary care EHRs. The authors selected variables by comparing relative frequencies of Read codes in the primary care case-related dataset compared to non-disease control, used the Random Forest Approach to minimise predictors and or associated variables, and used decision-making rules in the decision tree model. The authors deduced that ML approaches have the ability to determine the most informative predictors to accurately and reliably identify rheumatoid arthritis or other complex medical conditions in primary care EHRs [60]. Wu et al. have developed an automated asthma ascertainment algorithm from EHRs. The 10-fold cross-validation C4.5 decision tree algorithm was used to identify patients and to detect asthma status. The approach has shown that the ML algorithm is better than the F1-score rule-based approach, which has increased from 0.82 to 0.86 [61].

3) Logistic regression

Logistic regression is a supervised machine learning algorithm that was designed to learn classification problems. When the target variable is categorical, the problem is called a classification learning problem. The aim of logistic regression is to map a function from the dataset's features to the target classes in order to predict the probability that a new example belonging to one of the target classes [57]. Several researchers in the literature have used logistic regression methods. For instance, Weissler et al. constructed and validated a binary logistic regression to predict peripheral artery disease (PAD) using the least absolute shrinkage and selection operator (LASSO) method in the assigned patients. The authors concluded that their algorithm does not depend on clinical notes, and it can be used in cases where only administrative billing data (e.g., large administrative data sets) are available. A combination of diagnostic codes and administrative flags could reliably classify patients with PAD in large cohorts [62]. Jorge et al. randomly chosen patients with ≥ 1 systemic lupus erythematosus (SLE) ICD-9/10 codes from their EHRs, and they retrieved coded and narrative concepts from the training set using natural language processing and generated algorithms using penalized logistic regression to classify definite or defined/probable SLE. The authors concluded that their SLE algorithms worked well in internal and external validation [63].

2) Unsupervised machine learning

Unsupervised machine learning provides approaches to cluster EHRs on phenotypes or subtypes of patient groups. Unsupervised learning approaches can process very large data sets and do not require manual labelling, which makes it a method that is often used when designing high-throughput phenotyping systems [47]. Also, they may recognize patterns which collectively reflect original data in a compact and meaningful way, without any need for expert feedback or labelled examples [47] [56]. However, the validation of the outcomes for phenotypic groups using unsupervised learning strategies is difficult since there is no foundation for such groups. There are various unsupervised learning approaches have been developed including tensor factorization and clustering approaches [56].

1) Tensor factorization

A tensor, also known as a multiway array, is a higher-dimensional generalisation of a matrix (and a vector and a scalar). Tensor representations are useful because they can capture relationships in high-dimensional data [64]. Tensor factorization approach has been applied to EHRs. Kim et al. introduced a new approach for non-negative tensor factorization deriving from distinct and discriminative phenotypes. The authors interpreted and decomposed the co-occurrence of EHRs diagnoses and medications as third-order tensors with a computational phenotyping algorithm. They tested the discriminatory power of their models, and it showed superior results to state-of-the-art ICU mortality calculators [65]. Perros et al. used a PARAFAC2 tensor factorisation approach to derive clinically-meaningful phenotypes from medically complex children's longitudinal EHRs. The authors defined four medically complex phenotypes with different clinical characteristics among patients, and they concluded that PARAFAC2 can be used for unsupervised temporal phenotyping purposes when accurate definitions of different phenotypes are missing [66].

2) Clustering

The clustering methods is used to classify specific subgroups in a given dataset without providing a predefined hypothesis on the subgroups of properties [56]. Estiri et al. built an unsupervised clustering-based anomaly/outlier detection method to detect implausible observations in EHRs data. The authors tested different clustering approach specifications and computed confusion matrix indices against a collection of silver-standard plausibility thresholds, and found the clustering approach yielded results with exceptional specificity and high sensitivity [67]. Panahiazar et al. created a multidimensional patient similarity assessment technique that uses several types of EHRs information, and predicts a medication plan for each new patient based on previous related patient data. In their algorithm, patients were grouped into various clusters using a hierarchical clustering method, and a medication plan was subsequently allocated based on a similarity index to the overall patient population. Research findings indicate that it is reliable to use EHRs population-based data for an individual patient-specific assessment [68].

2.4.3 Combined approaches

Combined approaches typically incorporate both rules-based and machine learning strategies in order to improve algorithm sensitivity. Kagawa et al. created a novel framework for phenotyping combining both expert knowledge and a machine learning method and conducted binary classification to identify patients with Type 2 Diabetes Mellitus (T2DM). The authors found that their proposed framework can be used to build two types of phenotyping algorithms including one algorithm for screening, and another for identifying research subjects, and concluded that phenotyping algorithms based on their proposed framework are useful to extract T2DM patients in retrospective studies [69]. Jamian et al. designed and validated EHRs-based algorithms, including billing codes and clinical information to identify patients with systemic sclerosis (SSc) in EHRs. The authors conducted both rule-based and machine-based learning techniques for designing the algorithms, which have been calculated with positive predictive values (PPVs), sensitivities and F scores. They found that EHRs-based algorithms can classify SSc patients in a healthcare system that enables researchers to analyse significant results [70].

2.5 Validating and measuring the performance of electronic health records-based phenotyping

The validity of a phenotype definition is determined by its ability to accurately quantify or detect people that have and do not have the intended condition; that is, it must be able to accurately distinguish which individuals exhibit the true phenotype and who do not. The assessment of validity requires the use of a gold standard, which is described as the best standard available for determining the true or actual phenotype status [71]. Preparing a gold standard is a resource-intensive process that involves thorough manual analysis of clinical data [47]. Gold standard annotated corpora are essential resources for building and assessing the Natural Language Processing (NLP) systems. It is important to construct manually labelled instances that are applicable to specific NLP tasks. A useful gold standard should provide details, a wide range of documents, and annotated instances that reflect the diversity of document types and instances at stake in a particular task. This is important for 1) either train machine-learning NLP systems which require examples to learn from or find rules on rule-based algorithms and 2) evaluate NLP systems performance [72].

The performance of the EHRs phenotype algorithms could be measured using a number of metrics such as calculating positive predictive value (PPV), negative predictive value (NPV) sensitivity, specificity, F1-score and Area Under the Curve(AUC) as presented in Table 2.4 [47]. Pendergrass and Crawford defined the PPV as "the proportion of patients that have the disease identified by the algorithm and confirmed with manual chart review (true positives) among patients with the disease identified by the algorithm (true positives and false positives)", and they defined the NPV as "the proportion of patients that are classified by the algorithm as non-case or control and confirmed by manual chart review (true negatives) among all non-cases or controls identified by the algorithm (true negatives and false negatives)" [3].

Table 2.4: List of performance metrics commonly used to measure EHRs based-phenotyping approaches

Performance metrics	Definition	Equation
Sensitivity (Recall) (Se)	The proportion of all actually positive samples that are correctly detected.	$Se = \frac{TP}{TP+FN} \cdot 100$
Specificity (Sp)	The proportion of all actually negative samples that are correctly detected.	$Sp = \frac{TN}{TN+FP} \cdot 100$
Positive Predictive Value (precision) (PPV)	The proportion of positively detected samples that are true positive.	$PPV = \frac{TP}{TP+FP} \cdot 100$
Negative Predictive Value (NPV)	The proportion of negatively detected samples that are true negative.	$NPV = \frac{TN}{TN+FN} \cdot 100$
F1-score (F1)	The weighted harmonic mean of precision and recall.	$F1 = 2 \cdot \frac{Pre \cdot Re}{Pre+Re}$
Area under the ROC (AUROC)	ROC is the graph that represent the trade-off between sensitivity and specificity and area under the curve is equal the probability that a predictor will rank a randomly chosen positive sample higher than a randomly chosen negative one	

Source: (Alzoubi, 2019, P.15)

2.6 USES OF ELECTRONIC HEALTH RECORDS-BASED PHENOTYPING

2.6.1 Cross-sectional research

Phenotyping may be used for cross-sectional research to identify the relevant use cases include tracking adherence to recommendations for diagnosis and care, such as cervical cancer screening rates [73], and epidemiological studies, for example, assessing lifestyle, biological,

and environmental factors in addition to metabolic parameters to identify the factors associated with differences in metabolic health [74].

2.6.2 Cohort and case-control analyses

Phenotyping can also be used for cohort and case-control analyses on routinely collected EHRs data, for example, Suojalehto et al. conducted a retrospective, observational analysis to determine the characteristics of acrylate-induced Occupational Asthma (OA) in a wide variety of cases and compare others with OA induced by other Low Molecular Weight (LMW) [75], a study reported an increased risk of myocardial infarction in patients with rheumatoid arthritis treated anti-inflammatory medication rofecoxib [76], and the big expansion of available molecular phenotyping in large epidemiological cohorts offers diabetes research greater opportunity [77].

2.6.3 Translational research

Another important application of phenotyping is translational research, driving the evolving cross-disciplinary phenomics field [78] , for example, genomic data can now be linked to EHR-based phenotypes for more reproducible genome-wide association studies. Population-based repositories as well as other big cohorts provide adequate samples to detect new genetic associations with up to thousands of EHR phenotypes [79]. Phenotyping can also act as a framework for incorporating structured laboratory trials into clinical routine treatment, for example, electronic phenotyping can be used to assess clinical trial recruitment eligibility [80].

2.7 CHALLENGES ASSOCIATED WITH IDENTIFICATION OF EHRS BASED-PHENOTYPES

In spite of their many applications, the identification of phenotypes in EHRs is one of the most basic research challenges due to the heterogeneity, incompleteness and complex existence of EHRs data [81].

EHRs data are available as structured data, such as demographics, diagnostic codes, procedural codes, and laboratory results, combined with unstructured data, including progress notes, discharge summaries, and pathological reports[82][83]. Structured data can be easily analysed, stored and shared. However, it is difficult to retrieve embedded data from unstructured data and to make it computable and accessible. These data are shared irregularly over time and are usually fragmented throughout various institutions [49], and there is considerable variation in how data is entered into EHRs between providers and sites, partly as

a result of EHRs' multiple purpose as clinical records and billing software [84]. Additionally, the extraction of unstructured data from EHRs is time-consuming, costly and labour-intensive, for example, it needs involvement of clinical experts to analyse subsets in a patient record, and to validate algorithm performance, which takes considerable time [85][86].

Phenotyping requires the knowledge of some computer programming skills such as R packages to statistically analyse EHRs data in order to determine which patients have which diseases, for example, researchers and data scientists spend a significant amount of time using computer programs for: 1) identifying conditions of interest from diagnosis, treatments, and procedures, 2) creating phenotyping algorithms (such as diagnosis of rheumatoid arthritis and medication for rheumatoid arthritis), and developing specific inclusion and exclusion criteria [87].

Because identifying EHRs-based phenotypes is a time-consuming and labour-intensive process, reusing previously created EHRs-based phenotypes along with the used algorithms to conduct repeatable research becomes a compelling solution. Therefore, transparency in reporting of research methods in EHRs based research are needed to ensure that different researchers are using the same standards to identify patients.

2.8 CHALLENGES ASSOCIATED WITH REUSING OF EHRS BASED-PHENOTYPES

One of the most important factors for reproducible research is the availability of clinical codes in EHRs-based research because researchers, clinicians, and health informatics professionals often use them to identify the target population and their specific conditions, known as phenotyping [3] [14]. If researchers do not publish the code lists, they used (e.g., how they were established and the accurate phenotype definitions along with the original research using them), then an essential component of these studies is missing. In the absence of clinical code lists, data analysts would be unable to identify patients with or without conditions [14], and researchers would not be able to compare studies effectively. Even though code lists are available in some studies, researchers often encounter difficulties in retrieving relevant data from code lists created for another research project. Moreover, in specific uncommon conditions, minor errors in the selection of code lists may lead to misclassification of large numbers of patients, leading to biased results [88]. Although using previously developed phenotyping algorithms is often of interest to researchers in many studies, there are many challenges associated with reusing and replicating them effectively [89]. Therefore, it is extremely difficult to assess the validity and transparency of EHRs-driven studies [90].

Although researchers request better transparency in sharing clinical code lists [91] [92] , they face difficulties in obtaining comprehensive code lists from EHRs-based research. Although, there are currently no obligations from journals and funding parties to publish code lists, the Strengthening the Reporting of Observational Studies in Epidemiology and Reporting of Studies Conducted Using Observational Routinely Collected Health Data initiatives encourage transparency and open access to publicly available EHRs-based research [93] [94] [95].

2.9 CONCEPT LIBRARIES FOR EHRS BASED-PHENOTYPES

To address these various challenges, different data linkage centres in the UK and other countries, such as Canada, have developed data portals for phenotypes (concept libraries), such as the ClinicalCodes.org website [90], CALIBER research platform [6], and the MCHP Concept Dictionary and Glossary [96]. Building web-based concept libraries enables data analysts, researchers, and clinicians to upload and download lists of clinical codes, update previous code lists, and share clinical code data across platforms, which would improve the validation of EHRs-based research [90]. Concept libraries would contribute to the development of a group of users of EHRs data who are encouraged to share their methods. This will assist in maximising the benefits of EHRs data use and thus directly enhance health outcomes.

CHAPTER 3

METHODS

This chapter presents the various definitions of mixed methods research that have evolved over time; the rationale for selecting a mixed methods design; the four main characteristics of mixed methods research designs; mixed methods research design typologies; and the philosophical assumptions used in this thesis. It also describes in detail the exploratory sequential mixed approaches used in this thesis' three phases: 1) two qualitative studies; 2) the development of an email survey; and 3) one quantitative study, including data collection methods, design, application for ethical approval, and data analysis.

3 CHAPTER 3: MIXED METHODS RESEARCH

I used in this thesis a combination of qualitative and quantitative approaches, which is known as mixed methods research. Mixed methods research is becoming more common in health-related research because it allows for a deeper understanding of complex human phenomena [97].When diverse approaches are used, mixed methods research indicates that incorporating both qualitative and quantitative data results in a more comprehensive examination of the data. In addition, the use of mixed methods research provides complementarity (i.e., qualitative results may provide further insight into why a certain system may or may not be used successfully, as well as additional insight into the interpretation of quantitative results) [98].

3.1 DEFINITION OF MIXED METHODS RESEARCH

Many definitions of mixed methods research have developed over the years, each of which incorporates different aspects of methods, research processes, research purpose, and philosophy [99]. Johnson et al. offered a general definition for mixed methods research: *"Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration"* [100]. In this definition, the authors viewed mixed methods as a methodology that ranged from perspectives to inferences and included a combination of qualitative and quantitative research. Their goals for mixed methods "breadth and depth of understanding and corroboration" implied that they linked the definition of mixed methods to a justification for conducting it, and they suggested that a common definition should be used [99].

3.2 RATIONAL BEHIND CHOOSING MIXED METHODS DESIGN

I decided to employ a mixed methods design based on that the benefit of combining qualitative and quantitative data collection methods exceed using only one method to answer the research question [99]. Mixed methods designs are best suited to research problems where several perspectives on the research problem will provide a deeper understanding than a single perspective [101]. For example, research that simply collected a macro image of a health service utilizing quantitative data collection may overlook the factors that influence individuals accessing the service. The addition of a qualitative component that investigates the experiences of people who use the service would contribute greater depth to the understanding of the

problems [102]. Greene et al. identified five key purposes of mixed methods designs: triangulation, complementarity, development, initiation, and expansion, and they examined their justifications in depth [103]:

1. Triangulation: *"seeks convergence, corroboration, correspondence of results from the different methods to increase the validity of constructs and inquiry results by counteracting or maximizing the heterogeneity of irrelevant sources of variance attributable especially to inherent method bias but also to inquirer bias, bias of substantive theory, biases of inquiry context".*
2. Complementarity: *"seeks elaboration, enhancement, illustration, clarification of the results from one method with the results from another to increase the interpretability, meaningfulness, and validity of constructs and inquiry results by both capitalizing on inherent method strengths and counteracting inherent biases in methods and other sources".*
3. Development: *"seeks to use the results from one method to help develop or inform the other method, where development is broadly construed to include sampling and implementation, as well as measurement decisions to increase the validity of constructs and inquiry results by capitalizing on inherent method strengths".*
4. Initiation: *"seeks the discovery of paradox and contradiction, new perspectives of frameworks, the recasting of questions or results from one method with questions or results from the other method to increase the breadth and depth of inquiry results and interpretations by analysing them from the different perspectives of different methods and paradigms".*
5. Expansion: *"seeks to extend the breadth and range of enquiry by using different methods for different inquiry components to increase the scope of inquiry by selecting the methods most appropriate for multiple inquiry components".*

Bryman extended upon this scheme based on a comprehensive analysis of the various justifications for combining quantitative and qualitative research that are regularly presented in both methodological papers and research publications. The extended scheme provided the following justifications [104]:

1. Triangulation: *"refers to the traditional view that quantitative and qualitative research might be combined to triangulate findings in order that they may be mutually corroborated. If the term was used as a synonym for integrating quantitative and qualitative research, it was not coded as triangulation".*
2. Offset: *"refers to the suggestion that the research methods associated with both quantitative and qualitative research have their own strengths and weaknesses so that combining them allows the researcher to offset their weaknesses to draw on the strengths of both".*
3. Completeness: *"refers to the notion that the researcher can bring together a more comprehensive account of the area of enquiry in which he or she is interested if both quantitative and qualitative research are employed".*
4. Process: *"quantitative research provides an account of structures in social life, but qualitative research provides sense of process".*

5. Different research questions: *"this is the argument that quantitative and qualitative research can each answer different research questions, but this item was coded only if authors explicitly stated that they were doing this"*.
6. Explanation: *"one is used to help explain findings generated by the other'.*
7. Unexpected results: *'refers to the suggestion that quantitative and qualitative research can be fruitfully combined when one generates surprising results that can be understood by employing the other"*.
8. Instrument development: *"refers to contexts in which qualitative research is employed to develop questionnaire and scale items – for example, so that better wording or more comprehensive closed answers can be generated"*.
9. Sampling: *"refers to situations in which one approach is used to facilitate the sampling of respondents or cases"*.
10. Credibility: *"refers to suggestions that employing both approaches enhance the integrity of findings"*.
11. Context: *"refers to cases in which the combination is rationalized in terms of qualitative research providing contextual understanding coupled with either generalizable, externally valid findings or broad relationships among variables uncovered through a survey"*.
12. Illustration: *"refers to the use of qualitative data to illustrate quantitative findings, often referred to as putting 'meat on the bones' of 'dry' quantitative findings"*.
13. Utility or improving the usefulness of findings: *"refers to a suggestion, which is more likely to be prominent among articles with an applied focus, that combining the two approaches will be more useful to practitioners and others"*.
14. Confirm and discover: *"this entails using qualitative data to generate hypotheses and using quantitative research to test them within a single project"*.
15. Diversity of views: *"this includes two slightly different rationales – namely, combining researchers' and participants' perspectives through quantitative and qualitative research respectively, and uncovering relationships between variables through quantitative research while also revealing meanings among research participants through qualitative research"*.
16. Enhancement or building upon quantitative/qualitative findings: *"this entails a reference to making more of or augmenting either quantitative or qualitative findings by gathering data using a qualitative or quantitative research approach"*.

3.3 CHARACTERISTICS OF MIXED METHODS RESEARCH DESIGNS

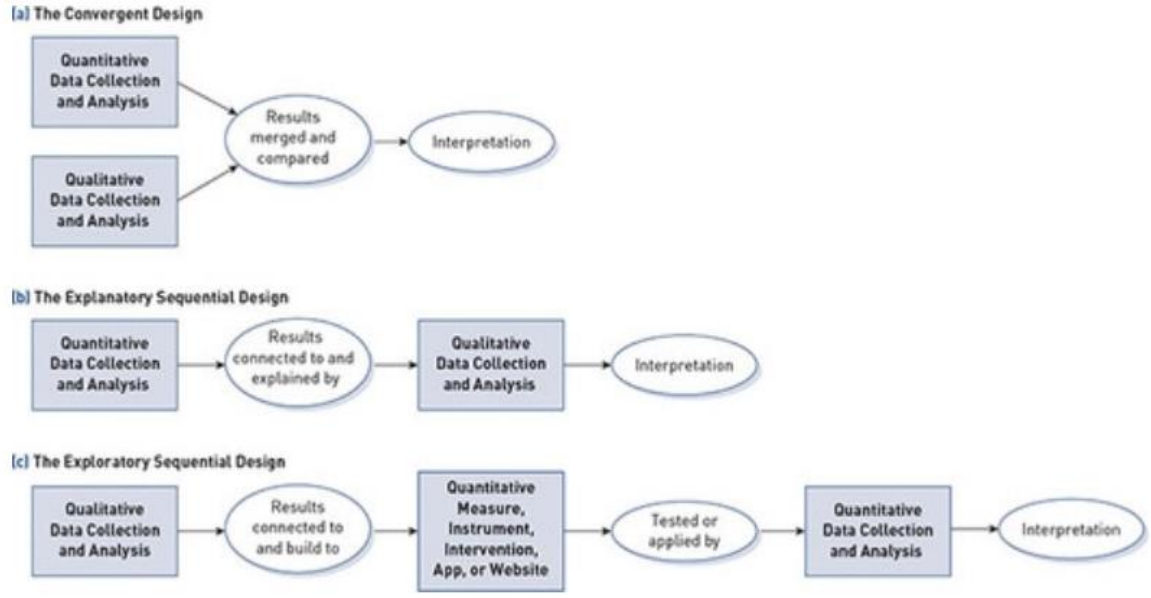
Mixed methods research designs are defined by four main characteristics: First, the level of interaction between qualitative and quantitative data (independence or dependence) [99]. This characteristic based on the extent to which qualitative and quantitative data will interact with one another or remain separate. Will one dataset collection, for example, inform the other? will the two datasets be acquired independently of each other? or will they be collected concurrently? [102] Second, the relative timing or sequencing of data collections implementation (simultaneous or sequential) [99]. Simultaneous / concurrent designs acquire both qualitative and quantitative data be collected at the same time. However, a sequential design involves collecting qualitative and quantitative data independently, using the outcomes from one method of data collection (e.g. interviews) as a basis for collecting another set of data

(e.g. survey) [102]. Third, the relative priority (or weighting) of the quantitative and qualitative parts in addressing the aim of the research [99]. Designs differ in the priority assigned to qualitative and quantitative data. Often, exploratory studies prioritize qualitative data. However, explanatory studies are likely to prioritize quantitative data [97]. Establishing the relative priority of each type of data before starting the study is particularly crucial if contradictory outcomes are revealed [102]. Fourth, the difference of the integration point of qualitative and quantitative data. Integration in a mixed methods research can occur anywhere in the research process [102]. A mixed methods research design should include at least one connection point of integration and may include more. In the explanatory and exploratory designs, each phase flows directly into another. However, the two types of data (qualitative and quantitative) are directly compared in the convergent designs [97].

3.4 TYPOLOGIES OF MIXED METHODS RESEARCH DESIGNS

Creswell et al. propose three core mixed methods designs as a useful framework for researchers planning their own research [99]. Through using a typology-based design, the researcher is given a framework and rationale to guide the application of the research methodologies, ensuring that the final design is rigorous and of high quality. As illustrated in Figure 3.1, the three core mixed methods designs are the convergent design, the explanatory sequential design, and the exploratory sequential design. Their various definitions, purposes, procedures, strengths, and challenges were described below.

Figure 3.1: General diagrams of the three core designs



Source: (Creswell & Plano Clark, 2017, P.84)

3.4.1 The convergent design

The convergent design was previously known as the concurrent or parallel design (Figure 3.1a) [99] [105]. It is a mixed methods design in which the researcher collects and analyses two separate data (quantitative and qualitative) at the same time and then merges the two data for the goal of comparing or combining the findings [99]. Purposes for using this design include demonstrating quantitative results with qualitative findings or conversely, and investigating correlations among variables by introducing new variables based on transformed qualitative data into the relationships [105]. The convergent design has four major steps [99] [105]. First, the researcher collects both quantitative and qualitative data on the subject of interest. These two methods of data collection occur concurrently but are usually distinct (i.e., one does not rely on the outcomes of the other), and they are usually equally important in answering the study's research questions. Second, the researcher uses quantitative and qualitative analytic procedures to analyse the two data sets separately and independently. Once the researcher has the two sets of preliminary results, he or she proceeds to the interface and works to merge the outcomes of the two data sets in the third step. This combining step may involve directly comparing the individual results.

This design has several strengths and advantages [99] [97] [105]: 1) it is an efficient design in which both types of data are collected at nearly the same time during one phase of the research making the population accessible, 2) this design is simple and straightforward. Therefore, it is frequently chosen by researchers who are new to mixed methods research, and 3) it allows for a direct comparison of participant perspectives acquired in an open-ended questioning format (e.g., semi-structured interview) with perspectives acquired from the researchers' perspective (e.g., using an instrument such as a survey selected by the researcher) in close-ended questioning. Despite its popularity in mixed methods, convergent design is perhaps the most difficult of the three core mixed methods designs. The following are some of the difficulties encountered by researchers who employ convergent design [99] [97] [105]: 1) it can be difficult to combine two sets of highly diverse data (e.g., one set of text and the other set of numbers) and their outcomes in a meaningful manner, 2) Because quantitative and qualitative data are typically collected for different purposes (i.e. quantitative generalisation vs. qualitative in-depth description), different sample sizes may occur. Therefore, when combining the two data sets, researchers must understand the implications of having different samples and different sample sizes, and 3) one of the most difficult aspects of a convergent design is knowing what to do when the findings diverge rather than converge.

3.4.2 The explanatory sequential design

The explanatory sequential design is a mixed methods design in which the researcher conducts a quantitative phase first and then follows up on specific outcomes with a qualitative phase to help explain the quantitative findings (Figure 3.1b) [97] [105]. The general purpose of this design is to employ a qualitative phase to provide a more detailed explanation of initial quantitative findings, and the name of the design (explanatory) reflects how the qualitative data helps in the explanation of the quantitative findings [99]. This design is most helpful when the researcher needs to evaluate trends and relationships using quantitative data while also explaining the reasons behind the findings [105]. The explanatory sequential design has four major steps [99] [105].The researcher begins by developing and implementing a quantitative phase, which comprises the collection and interpretation of quantitative data. In the second step, the researcher establishes a connection to a subsequent phase (i.e., the point of integration for mixing) by identifying quantitative findings that require additional explanation and using these findings to guide the development of the qualitative strand. As a result, the qualitative phase is dependent on and connected to the quantitative results. Finally, the researcher assesses how much and how well the qualitative data explain and contribute knowledge to the

quantitative data, and what generally is discovered in accordance with the purpose of the research.

The explanatory design has several advantages, including the following [97] [99] [105]: 1) its two-phase structure makes it simple to implement, as the researcher conducts the two methods (quantitative then qualitative) separately and collects just one type of data at a time, 2) the final result can be organised with a quantitative part followed by a qualitative part, making it simple to write and creating a clear separation for readers, and 3) this design enables emergent approaches in which the second phase can be developed according to the results of the initial quantitative phase. Although the explanatory design is simple, researchers who use this approach face the following challenges [97] [99] [105]: 1) This design needs a significant amount of time for the two phases to be implemented. While the qualitative phase may require limited number of participants, adequate time must be set aside for it, 2) the researcher must determine which quantitative results require additional explanation. While this cannot be exactly known until after the quantitative phase is complete, options such as identifying significant outcomes and strong predictors might be explored during the planning phase of the study, and 3) The researcher must decide who will be sampled in the second phase and what criteria will be used to select participants.

3.4.3 The exploratory sequential design

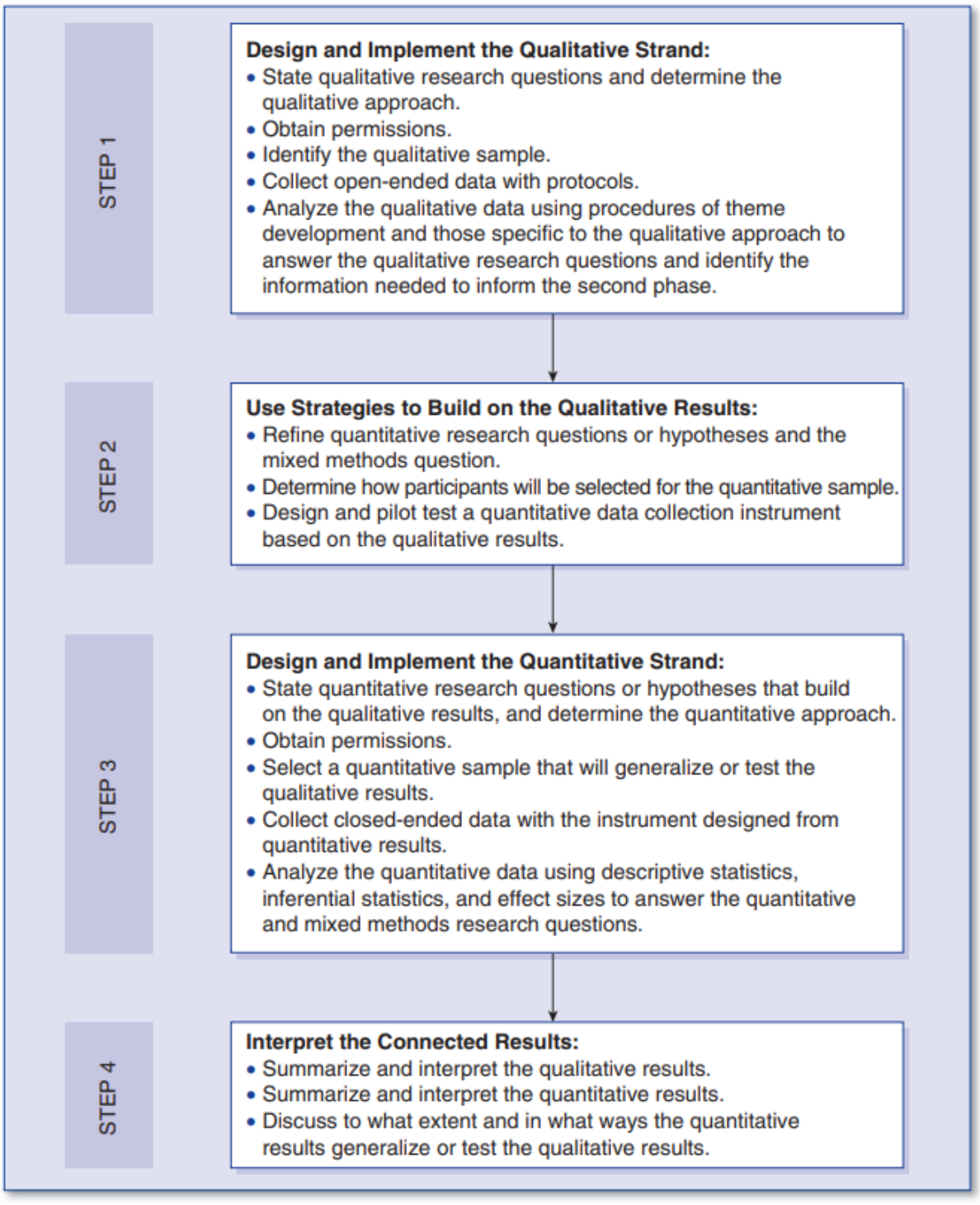
The exploratory sequential design, as shown in Figure 3.1c, is a three-phase mixed methods design in which the researcher begins with the collecting and analysing of qualitative data, followed by a development phase in which the qualitative findings are translated into a quantitatively tested technique or instrument [99] [105]. This indicates that the developed technique or instrument will be based on the perspectives of participants. The design name reflects the emphasis on exploration prior to the development phase [99]. The major purpose of the exploratory sequential design is to develop and implement a quantitative measure, survey, intervention, digital tool, or new variable based on qualitative data [105]. The intent of this design is to generate or inform the second quantitative method through the outcomes of the first qualitative method [101].

The exploratory sequential design has four major steps [99] [105]: it begins with the collecting and analysis of qualitative data to investigate a phenomenon. The researcher determines the outcomes on which the quantitative component will be developed in the next phase, which reflects the point of integration in mixing. In the development phase, the researcher creates an

instrument, identifies variables, or designs intervention tasks to test. These developments link the study's initial qualitative phase to the later quantitative strand. In the third step, the researcher conducts the quantitative strand of the study with a new sample of participants to examine the key variables using the developed instrument or intervention. Finally, the researcher evaluates the quantitative data in terms of how and to what extent they generalise or extend the initial qualitative findings. Figure 3.2 illustrates the basic procedures in implementing the exploratory design.

Many of the advantages of the exploratory sequential design are similar to those of the explanatory sequential design because in both designs just one type of data is collected at a time. Its distinct advantages include the following [97] [99] [105]: 1) because the exploratory sequential design is structured in phases, it is simple to describe, apply, and report, 2) although this design is typically qualitative in nature, integrating a quantitative component can assist quantitative-biased audiences in accepting the qualitative method, and 3) this design is useful when the need for a second quantitative phase becomes apparent as a result of the findings of the first qualitative phase. Despite all of these advantages, there are several disadvantages associated with using the exploratory sequential design, including [97] [99] [105]: 1) this design requires a significant amount of time to implement, possibly including time for a third phase to develop a feature (e.g., new instrument), 2) two different samples might need to be identified to improve the generalizability of quantitative data (i.e., researchers might consider utilising a small purposive sample in the first phase and a larger sample of different participants in the second phase, and 3) the researcher must decide what qualitative outcomes to employ to develop the quantitative feature and how to use these findings to develop quantitative measures or materials.

Figure 3.2: Flowchart of the basic procedures in implementing the exploratory design



Source: (Creswell & Plano Clark, 2017, P.98)

3.5 PHILOSOPHICAL FOUNDATIONS

In general, a philosophical assumption or worldview is the lens through which one observes the world [102]. In a mixed methods research, articulating philosophical assumptions involves acknowledging the worldview that serves as a foundation for the research, explaining the worldview elements and connecting these elements to specific procedures in the mixed methods research [99]. The four most commonly used worldviews in mixed methods research are the following:

1. Post positivism worldview:

This philosophy is the typical foundation for quantitative research that examines variables to establish inferences and generalisations about reality [99] [106]. When employed as the foundation for mixed methods research, it is common for researchers to prioritise quantitative methods, "quantitize" qualitative data for statistical analysis. Critics of this foundation for mixed methods research point out that it likely to restrict qualitative approaches to more structured approaches [106].

2. Constructivism worldview:

This philosophy is the typical foundation for qualitative research, works from a different set of assumptions. The understanding or meaning of phenomena, formed through participants and their subjective views, make up this worldview. When participants provide their understandings, they speak from meanings shaped by social interaction with others and from their own personal histories [99]. When utilised as the foundation for mixed methods research, researchers commonly prioritise qualitative methods, narrow the research to a small number of information-rich samples, and place emphasis on interpretations of each researcher [106]. Critics of this foundation for mixed methods research point out that it restricts quantitative methods to descriptive statistical analyses [106].

3. Transformative worldview:

This philosophical foundation is focused on the need for social justice and the pursuit of human rights [99]. It is positioned as a solid foundation for mixed methods research for those motivated by social justice concerns. When it is utilised as the foundation for mixed methods research, researchers typically prioritize the use of mixed methods within a theoretical framework to promote a social justice agenda by working for change for marginalised individuals [106]. Critics of this foundation for mixed methods research point out that it likely

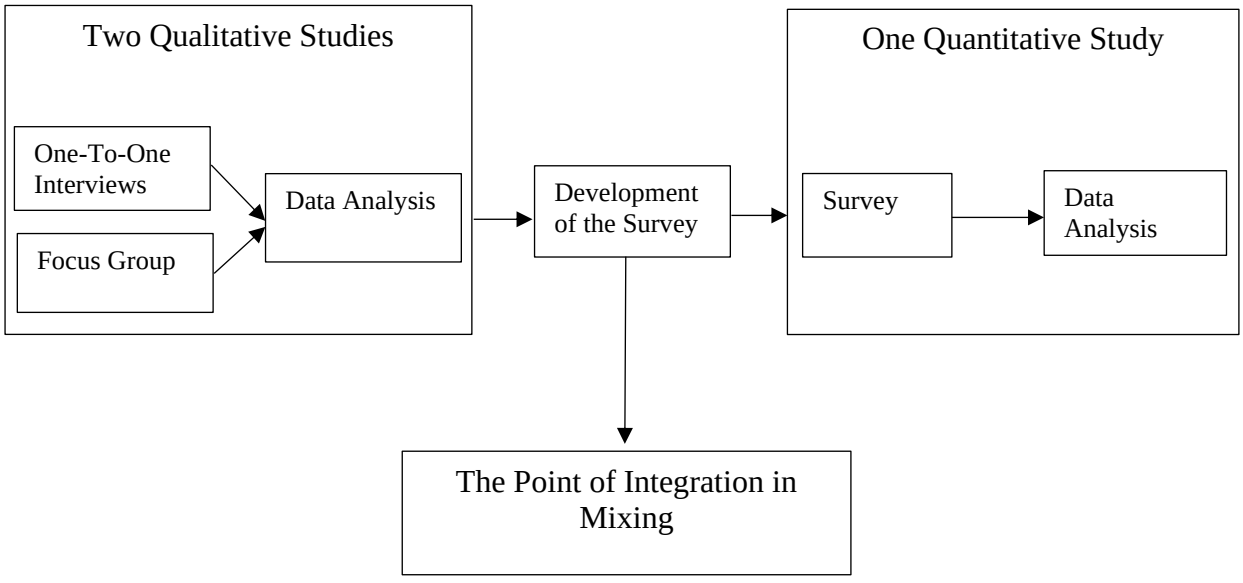
to overemphasise values and that its social justice agenda is properly viewed as a research purpose rather than a philosophical assumption [106].

4. Pragmatism worldview:

This philosophy is frequently promoted as a solid foundation for mixed methods research [106]. The emphasis is on the outcomes of research, on the value of the question addressed rather than the methods, and on the utilisation of multiple data collection methods to explain the issues under study. As a result, it is pluralistic and focused on "what works" and real-world application [99]. Critics of this foundation for mixed methods research point out that it inappropriately reduces the epistemic distinctions between quantitative and fail to recognize who decides what works [106].

In this thesis, I used the exploratory sequential mixed methods design to explore the needs of various users of concept libraries. To achieve this, I started by collecting and analysis of the qualitative data including one-to-one interviews and a focus group. In the next step, based on the findings, I identified variables and developed the quantitative component (i.e., a survey instrument), which reflects the point of integration in mixing. This development links the study's initial qualitative phase to the later quantitative strand. In the third step, this instrument is administered to a new sample of participants to examine the key variables. Finally, I analysed the quantitative data in terms of its generalisation or extension of the preliminary qualitative findings. Figure 3.3 illustrates the exploratory sequential mixed methods design used in this thesis.

Figure 3.3: The exploratory sequential mixed methods design used in this thesis



I chose the exploratory sequential mixed methods design for the following reasons: 1) the problem of my thesis is more qualitative in nature, and it makes sense to begin with a more inductive approach, 2) I needed to develop a survey instrument that is both substantively relevant and culturally sensitive (i.e., the survey elements were determined based on responses of the participants), 3) I had enough time to complete this thesis in three major phases (i.e., qualitative, development, and quantitative), 4) and I found new emerging research issues based on qualitative outcomes from a limited sample size that could be properly explored with a big quantitative sample. I applied many worldviews in this thesis, which is known as the pragmatism worldview, with the worldviews shifting from one phase to the next. During the first qualitative phase of this research (interviews and focus groups), I used constructivism worldview to appreciate different perspectives and gain a comprehensive knowledge. When I started the quantitative phase, the underlying assumptions evolved to a post positivism worldview, which directed the need for identifying and measuring variables and statistical trends.

3.6 THE FIRST PHASE: QUALITATIVE STUDIES USING INTERVIEWS AND FOCUS GROUP DISCUSSIONS

3.6.1 Data collection methods

Qualitative research provides comprehensive and interesting knowledge about the real worlds, experiences, and views of healthcare professionals and patients that is distinct but occasionally complimentary to the information obtained from quantitative research [107]. In this thesis, I conducted the following two qualitative studies:

1) The first qualitative research: one to one interviews with researchers, clinicians and managers were conducted to examine their specific needs of concept libraries.
2) The second qualitative research: a focus group with researchers working with SAIL databank was held for 2 hours, and a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system and the proposed concept library were performed.

3.6.2 Research ethics

The British Psychological Society Code of Human Research Ethics defined research ethics as "the moral principles guiding research from its inception through to completion and publication of results." [108]. Ethics should be considered at the planning stages and throughout the life of

a research project, especially if it contains primary research components such as surveys, focus groups, or interviews [109].

3.6.3 Core principles of research ethics

In this research, I considered the following core principles of research ethics, including: 1) social responsibility, 2) independence, 3) informed consent and voluntary participation, 4) anonymity and confidentiality, 5) integrity and transparency [108] [109].

1) Social responsibility: I am aware of my obligation to the communities in which I work in order to serve the public. My goal is to maximise the research's benefits while minimising the danger or harm to participants.
2) Independence: the research was conducted independently, and I declare that I have no conflicts of interest.
3) Informed consent and voluntary participation: the participants have provided adequate information to enable them to make an informed consent. Additionally, the goal, methods, and expected uses of the research, as well as the benefits associated with their participation, were explained orally during the interviews and focus group discussions, and in writing in the survey's email invitation. They were notified of their right to withdraw from the research at any time and for any reason without having to explain their decision. Participants were informed about how their data would be stored and handled.
4) Integrity and transparency: each stage of the research design, data collecting, coding, and analysis was thoroughly documented to ensure the research process was transparent.
5) Anonymity and confidentiality: The participants' anonymity and confidentiality were maintained throughout the study. To accomplish this, the participants' personal information was kept separate from their interviews, focus groups, and survey responses. Also, only my supervisors and I have access to the participants' personal information, and their sensitive personal information, such as names and work titles, was stored securely and protected by a password.

Additionally, I followed the eight Data Protection Principles, which provide the basic standard under the Data Protection Act 1998 [110]:

1. Personal data must be processed fairly and lawfully and shall not be processed unless certain conditions are satisfied.
2. Personal data must be obtained and processed for specific and lawful purposes. It must also only be processed for purposes which are compatible with those for which it was obtained.
3. Personal data must be adequate, relevant and not excessive for the purposes it is being processed for.
4. Personal data must be accurate and (where necessary) up to date.
5. Personal data must not be kept for longer than is necessary.
6. Personal data shall be processed in accordance with the rights of data subjects.
7. Personal data shall be protected by appropriate technical and organisational measures against unauthorised or unlawful processing and against accidental loss, destruction or damage.
8. Personal data shall not be transferred outside the European Economic Area without adequate protection for the rights and freedoms of data subjects in relation to personal data.

3.6.4 Application for ethical approval

The application for ethical approval to conduct the interviews and focus group discussions was submitted to the SUMS Research Ethics Sub-Committee on January 31, 2019 (Appendix 3), and the approval letter was obtained from the SUMS Research Ethics Sub-Committee on March 7, 2019 (Appendix 4).

3.6.5 The first qualitative research: one to one interviews

Interviews are an essential data collection method that includes verbal conversation between the researcher and the participant. In qualitative research, interviews are commonly used as a data collection method [111]. For example, they are widely used in survey designs and in exploratory and descriptive research [112]. Interviews are often used as a research approach to collect information on participants' experiences, perspectives, and beliefs regarding a particular research issue or phenomenon [113].

There are three main of interviews types: structured, semi-structured and unstructured [112]. The most rigid type is the structured interview, in which the interviewer asks the same questions to all participants in the same order. A closely structured collection of questions,

similar to a questionnaire, is used. The questionnaire's questions must be prepared in advance, preferably with the assistance of a pilot study to clarify the questions [112]. Multiple interviewers can conduct interviews simultaneously because everyone asks the same questions in the same way, allowing for interviews with a bigger set of participants [114].

In an unstructured interview, the aim of the interviewer is to address a small number of issues, usually one or two, and asks subsequent questions based on the interviewee's previous response [112] [114]. Although only a few issues are addressed, they are well explored. These are interviews when the interviewer needs to learn about a given issue but has no framework, predetermined plan, or expectation of how the interview will go [112].The questions may differ among interviews, based on the responses of the participants. Typically, the same interviewer conducts the interview with all participants [114].

The semi-structured interview enables the interviewer to use a relatively consistent collection of open-ended questions to ask each participant in roughly the same order [112] [114]. However, the format of this interview enables the interviewer to ask clarifying questions as needed and follow the participant's thoughts [114]. The open-ended type of the questions specifies the study's subject while allowing both interviewer and interviewee to explore further into specific issues. In a semi-structured interview, the interviewer is free to probe the interviewee for clarification on the initial response or to follow an interviewee-initiated line of questioning [112]. Semi-structured interviews are useful for collecting large amounts of viewpoint data or when the research is exploratory in nature and it is not possible to create a list of probable pre-codes due to a lack of knowledge about the subject area [112].

3.6.5.1 Design

For the purpose of this research, I developed semi-structured interview questions to use them in one-to-one interviews (presented in Table 3.1). I chose to conduct in-depth one-to-one interviews because they allow collecting detailed information from a small number of participants and providing understanding into their variety of thoughts, needs, and experiences regarding concept libraries. In-depth interviews are qualitative research methods including intensive one-to-one or individual interviews with a small number of participants [115]. Individual interview is a common data collection approach in health and social research [111]. Although one-to-one interviews are labour intensive, they might be the most effective method of collecting high-quality data. In comparison to other data collection methods, one-to-one interviews allow for greater flexibility, and they can help to in-depth data collecting by

providing insight into the participants' opinions, understandings, and experiences of a given issue [111] [112].

3.6.5.2 Data Collection

I employed purposeful sampling, so I established criteria for selecting participants based on the aims of the qualitative study (one-on-one interviews), which included selecting people from all potential disciplines that work with linked routine data and disease phenotyping. I contacted the SAIL Databank coordinator about the disciplines of all SAIL Databank users, and then I identified people who fit these criteria.

Six participants from a variety of disciplines, including researchers (3/6, 50%), a clinician (1/6, 17%), a machine learning expert (1/6, 17%), and a senior research manager (1/6, 17%), at Swansea University and Cardiff University, were asked to participate in one-to-one interviews by email. The invitation email specified the aim and purpose of this study, the duration of each interview (30 minutes), and the location of the interviews, which might be their offices or a convenient and private location on the Swansea University campus.

Semi-structured interview questions, which follow the structure proposed by Krueger and Casey [116], were used (Table 3.1). The structure of the interview questions consisted of introductory, flow, key, and final questions. The purpose of the introductory questions was to help the participants talk freely about their overall experiences. The flow questions were designed to create a smooth transition to the key areas that the authors intended to explore. The final questions were designed to summarize the interview and ensure that the participants did not have further comments [117].

Before conducting the one-to-one interviews, I explained the purpose of the research and what it involved, and at the beginning of each interview, participants received additional verbal and written information about the research project, including a consent form (Appendix 1) and a participant information sheet (Appendix 2). The interviews were conducted at Swansea University Medical School in a place selected by the participants (e.g., their office). After 5 interviews, no new themes were observed and interview 6 confirmed that no new themes emerged. The interviews were audio recorded and transcribed verbatim. Thematic analyses were then performed using the 6 steps of Braun and Clarke to identify the themes and subthemes [118].

Table 3.1: One to one interview questions guide

Introductory questions	**Flow questions**	**Key questions**	**Final questions**
To improve repeatable research in Swansea, a team of developers is developing a prototype concept library. This is a portal that allows access to the Read codes or International Classification of Diseases–10 codes to identify conditions. Do you think this will be a helpful resource? Is the concept library a good idea that we should continue to develop?	Do you know about other already existing concept libraries? What do you think about them? Something like this exists at UCL called CALIBER. Have you seen CALIBER? Have you used it?	Do you prefer to use ready-made algorithms or to have access to them to modify them? In your opinion, how should codes and algorithms be validated, and should they be validated? (Why should or should not?) There are often different versions of a diagnosis (e.g., highly specific and suspected or likely cases). Do you think we need to collect and validate the best two versions of a diagnosis (specific or suspected)? Or do you think we should put all possible methods of identifying a condition, valid or not, and allow the researcher to choose?	What are your requirements for the concept library for it to be helpful and user-friendly? What developments would you like to see to improve repeatable research using routine data?

3.6.6 The second qualitative research: focus group discussions

Focus group discussion is one of the most common qualitative data collection method used in exploratory research that involves focusing on a particular topic, with a predefined group of people participating in an interactive conversation [119] [120]. It is designed to collect data from a purposefully chosen group of participants based on the primary goal of the research rather than from a statistically representative group of a larger population [119]. The primary objective of focus group research is to elicit a range of viewpoints on a topic of interest and to obtain a knowledge of the issues from the participants' perspective [120]. The focus group discussions can be conducted in a variety of approaches, including unstructured, structured, or semi-structured. [121].

In this research, I decided to adopt a semi- structured approach to conduct the focus group to achieve the following purposes:1) to explore the viewpoints of various users, such as researchers, health professionals, clinicians, and designers, such as health informatics teams, in developing a concept library (a portal of definitions for disease phenotyping), about which little is known and where their diverse needs are not addressed in the literature, 2) to collect a diverse range of experiences and perspectives through a collaborative discussion that cannot be gathered by individual interviews, 3) to identify the reasons why existing concept libraries are underutilised, and 4) to design a concept library usability survey based on the focus group findings, which will assist in determining the components to include.

Because I chose to adopt a semi-structured approach, I created a list of ten questions based on the objectives of this research, before to the focus group session. The purpose of the questions was to generate thoughtful and thorough responses from the participants; therefore, I avoided using closed-ended questions (e.g., yes or no). I used the semi- structured approach because it is the more flexible approach for this research, which included two focus group discussions. Although the two moderators (my first supervisor and I) used the same list of questions, the sequence of questions was adjustable to the needs of each group. Also, comparing of responses could be achieved since each focus group exposed to same list of questions.

3.6.6.1 Data Collection

All researchers working with the SAIL databank, a national e-health data linkage infrastructure in Wales (N = 34), were invited by email to participate in the focus group. Those researchers were chosen since it was very convenient for them to attend the focus group, which was held

at Swansea University in Wales. Because the SAIL databank is part of the broader UK data linkage repositories, I think the conclusions of this sample could have implications for other UK data linkage repositories. Only 14 (14/34, 41%) researchers attended the focus group discussions. (Table 3.2 shows general information about the participants).

In total, two focus group discussions, each of which had 7 (7/14, 50%) participants, were held for 2 hours by two moderators (my first supervisor and I) using the same set of semi-structured questions to perform a SWOT (strengths, weaknesses, opportunities, and threats) analysis for the current system for phenotyping and the proposed concept library. We used a SWOT analysis tool in this study because it enabled the participants to discuss what they liked (strengths), what advantages would be gained (opportunities), and what problems (weaknesses) and issues (threats) they felt needed to be tackled. Although the two moderators used the same set of questions, the order of the questions was adjusted to the needs of each group.

Table 3.2 A summary of general information on the participants in the focus group discussions (n=14)

Parameters	Information
Current job position, n (%)	Data scientist, 13 (93) Financial planner, 1 (7)
Sex, n (%)	Female, 5 (36) Male, 9 (64)
Education, n (%)	PhD degree, 6 (43) Master's degree, 6 (43) Bachelor's degree, 2 (14)
Research interests	Data Scientists • Concept Libraries • Repeatable Research with large health data • Phenotyping and Code lists of Cancer Disease • Respiratory Disease • Algorithm/ Reusable codes development • Asthma • Collaboration in research methods • Data Analysis • Machine learning • Arthritis • Health informatics • Musculoskeletal • Healthy aging • Gut – Brain Axis • Neurodegenerative conditions

- Statistical Methods
- Epidemiology
- Cancer

Financial Planners

- Intervention between primary care and secondary care and how they interact

Two focus group discussions (each consisting of seven participants) were held for 2 hours by two moderators (my first supervisor and I) using the same list of questions, and a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system and the proposed concept library were performed.

3.6.6.2 SWOT Analysis: Definition, Advantages and Limitations

SWOT analysis is a famous four-box strategy analysis and development methodology. The term SWOT stands for Strengths, Weaknesses, Opportunities, and Threats [122]. SWOT analysis has been used for decades and has the potential to become the most commonly used strategy tool. It is utilized by health and education, business, as well as charitable organisations [122] [123]. It is a useful planning framework that is used to evaluate an organisation, a strategy or a project. SWOT analysis considers all elements affecting the organization's success, both positive and negative. It has two dimensions: internal and external [124]. Internal dimensions include organisational factors (i.e., strengths and weaknesses). Whereas external dimensions include environmental elements (i.e., opportunities and threats). Strengths and opportunities are beneficial in achieving organisational goals, whereas weaknesses and threats are harmful in accomplishing organisational goals [124] [125].

Analysis of internal dimensions is used to determine the organization's resources, capabilities, core competencies, and competitive advantages. However, analysis of external dimensions identifies potential threats and opportunities by investigating resources of competitors, the industry environment, and the overall environment [126].

SWOT Analysis reveals existing status of an organization and allows for the development of future action plans. When applied correctly, the technique can provide a good base for strategy design. Despite the fact that it is a basic managerial tool with many benefits in the planning process, it also has limitations. The qualitative study of internal and external dimensions can only serve as a starting point for a more in-depth analysis during the planning phase [124]. I used a SWOT analysis tool in the early phases of this thesis as a foundation to development of quantitative tool (i.e., designing a concept library usability survey). Since academic studies on

the subject imply that combining qualitative and quantitative methodologies can improve the effectiveness of SWOT analysis [124], I will use a quantitative methodology in the second phase of my thesis.

At the beginning of the focus group discussion, I gave a brief presentation about concept libraries, including defining concept libraries, explaining their potential uses, and mentioning examples of some of the existing concept libraries in the United Kingdom. Swansea University developed a concept library prototype, which is software that lets users explore and create lists using a familiar web paradigm. Lists can be created through identifying codes using various methods such as keyword searches, regular expressions, and more complex rules within the SAIL Databank [127]. I asked one of the Swansea University prototype concept library developers to give a second presentation about it. Feedback from the participants was sought concerning their perceptions of the concept library's needs and their evaluation of the strengths and limitations of the proposed concept library. Participants' perceptions of existing concept libraries, as well as their assessment of the proposed concept library's strengths and limitations, were explored using the following set of semi-structured questions:

- What are your thoughts regarding the proposed data portal for phenotypes (a concept library) when it rolls out?
- Do you think this is worth doing? Would you value this?
- Has anybody used existing concept libraries? What have you experienced with them?

Let us talk now about your current system for phenotyping:

- What do you do? What are your methods?
- Are you happy with them? Or what would you like differently?
- What are your thoughts on this plan (building a concept library)?
- Would you use it? Would you share your phenotypes and your phenotyping algorithms?

If you do not want to share your work:

- Can you tell us why? And what motivates you to share it with others?
- Of all the things we have discussed, what is most important to you?

- Is there anything we should have talked about but did not?

The goal of using a SWOT analysis was to identify positive factors that operate together and the potential difficulties that must be identified and solved, and to allow participants to make a decision and express their perspective on the four factors: strengths, weaknesses, opportunities, and threats, in order to enhance the collective perception [122]. During the focus group discussions, participants expressed their own opinions and listened to the opinions of others. As the discussions progressed, participants began to ask questions of one another and share similar experiences. This increased the depth of the conversation. The SWOT analysis gave us a full picture of views and experiences of concept libraries by the participants, making this a holistic evaluation with the ability for participants to hear and comment on each other's responses.

3.6.7 Data analysis

Thematic analysis is an effective and efficient method for understanding a collection of experiences, thoughts, or behaviours among a set of data [128]. Braun and Clarke defined thematic analysis as "a method for systematically identifying, organizing, and offering insight into patterns of meaning (themes) across a data set" [128]. Guest et al. mentioned that thematic analysis focuses on detecting and defining both implicit and explicit ideas within the data (themes), codes are then often constructed to reflect the discovered themes, and they are used or connected to raw data as summary identifiers for subsequent analysis [129]. Thematic analysis requires interpretation during the selection of codes and the development of themes [130].

Thematic analysis is flexible enough to be used in various theoretical and epistemological contexts, as well as to a number of study questions, designs, and sample sizes [130]. Braun and Clarke mentioned that thematic analysis is not paradigm-specific; rather, it can be utilised in both realist/essentialist and constructionist paradigms [118]. Using thematic analysis in various research paradigms means applying this method to different goals and outcomes [130]. The following benefits of thematic analysis were outlined by Braun and Clarke [118]:

"1) relatively easy and quick method to learn and do, 2) results are generally accessible to educated general public, 3) useful method for working within participatory research paradigm, with participants as collaborators, 4) can usefully summarise key features of a large body of data, and/or offer a 'thick description' of the data set, 5) can highlight similarities and differences across the data set, 6) can generate unanticipated insights, 7) allows for social as

well as psychological interpretations of data, and 8) can be useful for producing qualitative analyses suited to informing policy development".

In this thesis, I utilized the following six phases of thematic analysis provided by Braun and Clarke [118] [128]:

1. Self-Familiarizing with the data: *"Transcribing data (if necessary), reading and rereading the data, noting down initial ideas".*

2. Creating initial codes: *"Coding interesting features of the data in a systematic fashion across the entire data set, collating data relevant to each code".*

3. Searching for themes: *"Collating codes into potential themes, gathering all data relevant to each potential theme".*

4. Revising themes: *"Checking in the themes work in relation to the coded extracts (Level 1) and the entire data set (Level 2), generating a thematic 'map' of the analysis".*

5. Identifying and naming themes: *"Ongoing analysis to refine the specifics of each theme, and the overall story the analysis tells; generating clear definitions and names for each theme".*

6. Writing up the report: *"The final opportunity for analysis. Selection of vivid, compelling extract examples, final analysis of selected extracts, relating back of the analysis to the research question and literature, producing a scholarly report of the analysis".*

The interviews and the focus group discussions were analysed separately following Braun and Clarke's (2006) six thematic analytic steps [118] [131]. The transcripts of the interviews and the focus group discussions were read several times and then initial codes were grouped into themes and subthemes using a qualitative data analysis software (NVivo) [132]. During the coding process, each data item received equal consideration to ensure that the codes were comprehensive.

I read all the transcripts, and my first supervisor read a sample of the transcripts. We independently identified the themes and subthemes, then met regularly to compare them and to reach an agreement on what was being done. Themes and subthemes were discussed concerning their relevance to the research question in the data collected. We critically reviewed themes again to determine their primary meanings, and similar initial themes were joined into one theme. We discussed the definitions of the relevant themes to the research questions and

applied appropriate names to describe each theme in this research. To ensure that all steps of the analysis are completed adequately, sufficient time has been allocated to avoid speeding through any of them. See table 3.3 for a further description of the thematic analytic steps used for this research.

Table 3.3 The 6 thematic analytic steps used for this research

Thematic analytic steps	
1. Self-Familiarizing with the data I transcribed half of the audio recordings from the interviews (n=3). The other half of the audio recordings from the interviews (n=3) and the audio recordings from the focus group discussions were transcribed by professional transcribers. During this phase, I read all of the interview and focus group transcripts several times, and my first supervisor read samples of them. We considered all the topics discussed by the participants, recorded notes on these topics in the transcripts, and then organised them in a note book.	**2. Creating initial codes** After we familiarised ourselves with the data, we worked independently to identify initial codes from the transcripts that summarized what was said during the interviews and focus group discussions. We organised the identified codes into meaningful groups using qualitative data analysis software (NVivo, QSR International). We used the same coding procedure for all the transcripts.
3. Searching for themes We started interpreting the initial codes using their extracted data, and we began grouping the codes with similar meanings together. Then, using the NVivo software (QSR International), the initial codes were then sorted and labelled into themes and subthemes depending on the meaning or relations shared by the codes.	**4. Revising themes** We critically reviewed and refined themes against the data several times to determine their core meanings, and similar initial themes were combined into one theme. To reach an agreement, themes and subthemes were discussed in terms of their relevance to the research question.

5. Defining themes

Each of the themes identified in the previous steps was named and defined. We used the initial labels created for the themes to provide appropriate names that describe the meaning of the themes in this study. We defined each theme based on the content and meaning of their codes, and we examined these definitions in relation to their relevance to the research questions.

6. Writing up the report

After defining and naming the themes, we began writing the findings for this study. We used quotes from the participants' responses that related to the themes and the research question to illustrate the findings.

3.7 THE SECOND PHASE: DEVELOPMENT OF THE CONCEPT LIBRARY USABILITY SCALE

An e-mail survey instrument, the Concept Library Usability Scale, was developed based on the findings of the first qualitative phase of this research. This development links the study's first qualitative phase to the subsequent quantitative phase, which represents the point where the two phases mix. Technological advancements such as the invention of the internet have had a major effect on the world as a whole over the last 50 years [133]. This effect is seen in the way survey research is conducted in the current period (i.e., quantitative online surveying). Lately, with the appearance of the Covid 19 pandemic, the world moves toward online surveying as the primary means of collecting survey data [134]. Quantitative online surveys can be used when researchers want to answer a question about large groups of people and/or generalise the results of the survey [133]. The majority of questions in quantitative online surveys are closed-ended. This means that the respondent is limited to a specific number of possible responses and must select just one of them. Closed-ended questions are also referred to as quantitative questions since the availability of response possibilities allows the researcher to turn the responses into numerical values, which makes statistical analysis easier [135].

There are three types of online surveys: e-mail surveys, website surveys, and smartphone surveys. E-mail surveys are inexpensive to produce and distribute, and are one of the most popular online surveys since anyone with access to online survey software, such as SurveyMonkey, Zoomerang, or Instant Survey, could construct an e-mail survey [136].

3.7.1 The e-mail surveys:

Sue and Ritter identified the following are specific advantages of the e-mail surveys [136]:

1) speed: An e-mail questionnaire can be sent to hundreds or thousands of people by entering or importing a distribution list and hitting the send button. Responses typically are received quickly, and data can be described and distributed via the software tool in real time.

2) economy: most e-mail software vendors (such as those mentioned earlier) offer free versions of their services. The free software often limits the number and types of questions and responses allowed. If these limitations pose a problem, a low-cost, monthly contract may be purchased that will expand the options and offer the survey creator the vendor's full suite of tools.

3) convenience: online survey software allows researchers to create the questionnaire, write the e-mail invitation, upload a distribution list, and send reminders directly from the software. In most cases, it is a seamless approach that automatically inserts such elements as the survey link and a link for respondents to opt out of the survey if they so choose.

4) simplicity: online survey software of the type we have been referencing does not require technical expertise on the part of the survey developer. Tools such as SurveyMonkey and Zoomerang are user-friendly, offer a selection of survey templates to jump-start the questionnaire creation process, and contain help features that include step-by-step instructions, tutorials, and online chats with support staff.

In this phase, I used the themes from the initial qualitative phase (both the interviews and the focus group) to identify existing concept library usability scales or surveys, but none were available. Therefore, I reviewed some of the existing standardized usability surveys to decide whether to use, merge, or modify some them according to this research's questions. A variety of usability surveys are widely available and frequently used, and are often consist of a specific set of questions provided in a specific order and using a specific format, with defined guidelines for generating scores based on the responses of participants [137] [138]. The following are two of the most well-known:

1) System Usability Scale (SUS). The survey designed by John Brooke at Digital Equipment Corporation, consists of ten Likert-type questions with responses on a 5-point scale [139].

2) Computer System Usability Questionnaire (CSUQ). The survey was created by James Lewis at IBM and consists of 19 questions on a 7-point scale [140].

3.7.2 System Usability Scale (SUS):

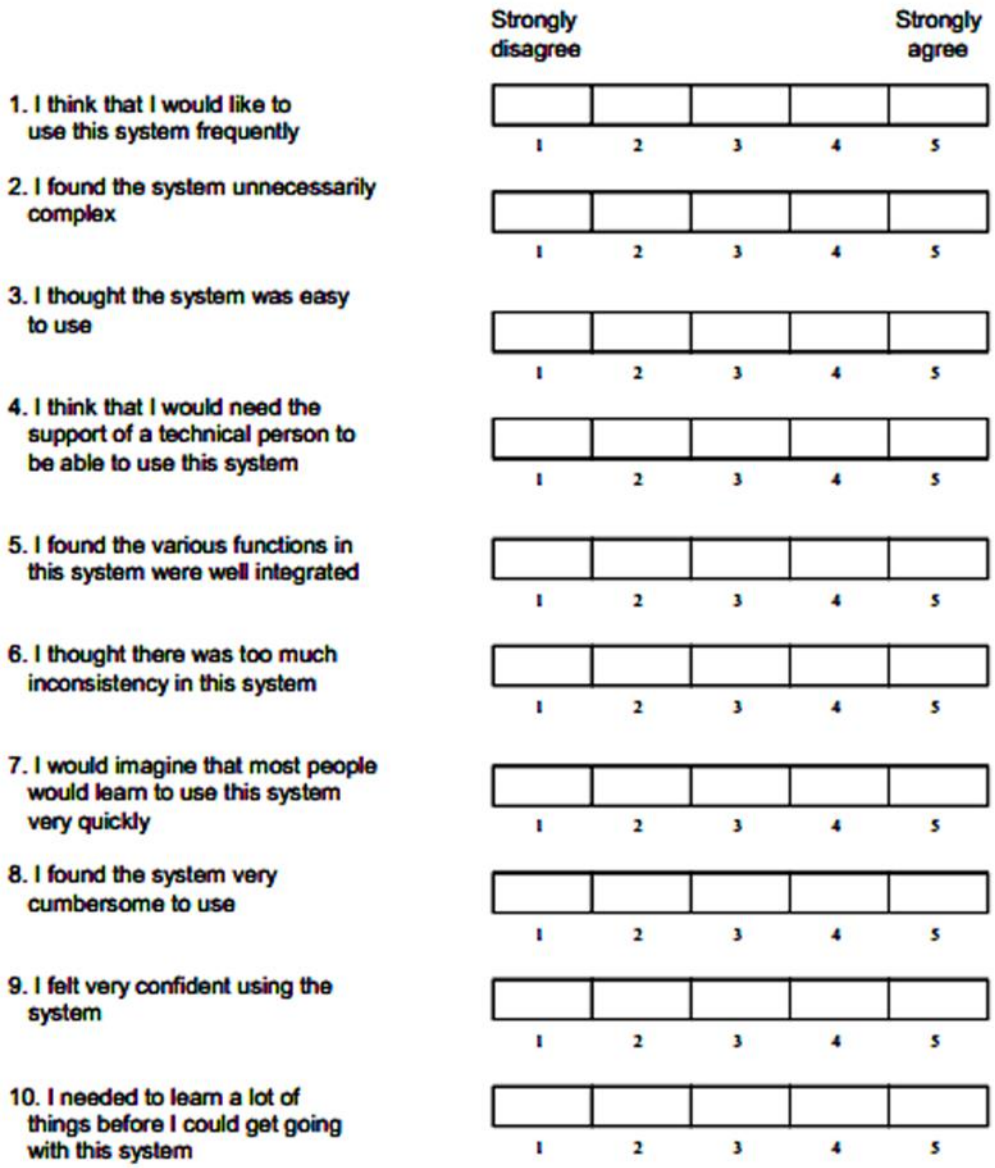
The System Usability Scale (SUS) was created by John Brooke in 1986. The SUS is a reliable (Likert scale) instrument for testing usability. SUS enables evaluation of a wide range of products, such as hardware, software, mobile devices, websites, and applications. SUS

consists of the following10-item survey with five response categories, ranging from strongly disagree (1) to strongly agree (5) [139]:

1. *I think that I would like to use this system frequently.*
2. *I found the system unnecessarily complex.*
3. *I thought the system was easy to use.*
4. *I think that I would need the support of a technical person to be able to use this system.*
5. *I found the various functions in this system were well integrated.*
6. *I thought there was too much inconsistency in this system.*
7. *I would imagine that most people would learn to use this system very quickly.*
8. *I found the system very cumbersome to use.*
9. *I felt very confident using the system.*
10. *I needed to learn a lot of things before I could get going with this system.*

Figure 3.6 represents the ten statements in SUS. The System Usability Scale is widely used because the word system is used in the statements, which reflects its initial use for software evaluation, could be replaced with a website, product, or interface without impacting the outcomes [137]. In addition, Brooke mentioned that "the selected statements actually cover a variety of aspects of system usability, such as the need for support, training, and complexity, and thus have a high level of face validity for measuring usability of a system" [139].

Figure 3.4: System Usability Scale Questionnaire



Source: (Brooke,1986, P.4)

3.7.3 Computer System Usability Questionnaire (CSUQ)

The Computer System Usability Questionnaire (CSUQ) is slightly longer than the SUS but is still manageable. The original statements, like the SUS questionnaire, use the word system. The CSUQ consists of the following 19 questions and uses a seven-point scale, ranging from strongly agree (1) to strongly disagree (7) [140]:

1. *Overall, I am satisfied with how easy it is to use this system.*
2. *It was simple to use this system.*

3. *I could effectively complete the tasks and scenarios using this system.*
4. *I was able to complete the tasks and scenarios quickly using this system.*
5. *I was able to efficiently complete the tasks and scenarios using this system.*
6. *I felt comfortable using this system.*
7. *It was easy to learn to use this system.*
8. *I believe I could become productive quickly using this system.*
9. *The system gave error messages that clearly told me how to fix problems.*
10. *Whenever I made a mistake using the system, I could recover easily and quickly.*
11. *The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.*
12. *It was easy to find the information I needed.*
13. *The information provided for the system was easy to understand.*
14. *The information was effective in helping me complete the tasks and scenarios.*
15. *The organization of information on the system screens was clear.*
16. *The interface of this system was pleasant.*
17. *I liked using the interface of this system.*
18. *This system has all the functions and capabilities I expect it to have.*
19. *Overall, I am satisfied with this system.*

3.7.4 The Concept Library Usability Scale

Based on the findings of the first phase of this research, a requirement for a subjective concept library usability scale/survey was recognised. Therefore, I created an e-mail survey instrument, which I named the Concept Library Usability Scale, using SurveyMonkey software. This development links the study's initial qualitative phase to the later quantitative phase, which reflects the point of integration in mixing. I aimed for the survey to be quick and easy to administer while also being reliable enough to be used to assess the user experience of a concept library. Several versions of the e-mail survey have been substantially reviewed with my supervisors, and we finally decided on the final version after ensuring that the questions are clear and comprehensible. Figure 3.7 presents the final version of the e-mail survey instrument, the Concept Library Usability Scale. The Concept Library Usability Scale contains 12 statements, and participants choose one of five response options ranging from "strongly disagree" to "strongly agree" to express their level of agreement or disagreement with the first ten statements. The eleventh statement was intended to allow participants to openly express and discuss their thoughts, while the twelfth statement invited participants to submit contact information if they were interested in participating in a one-to-one interview.

The following are the first ten items on the Concept Library Usability Scale:

1. I think that I would like to use this concept library frequently.

2. I think that I would need the support of a technical person to be able to use this concept library.

3. I found the various functions in this concept library, such as searching and viewing concepts; and creating and editing concepts, were easy to use.

4. I felt very confident using the concept library.

5. I needed to learn a lot of things before I could get going with this concept library.

6. I think that the user documentation is task oriented and consists of clear, step by step instructions.

7. I found the concept library supports advanced functional tasks (e.g., it allows using of programming languages such as R, SQL, or Python).

8. I feel it is acceptable if I am required to reference the concept library when publishing papers.

9. I found that it was easy to understand how the concept library is run and managed.

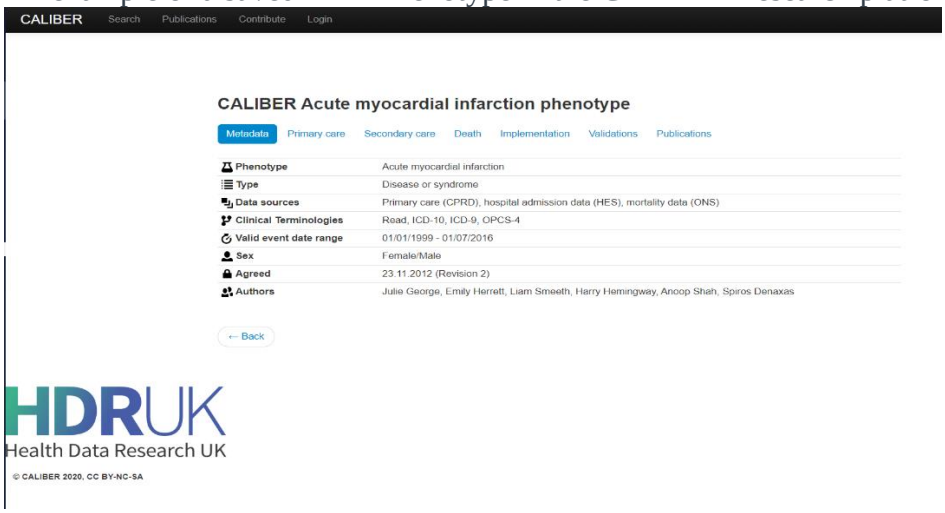
10. I thought the concept library supports clear algorithms labelling convention.

Figure 3.5: The Concept Library Usability Scale

Swansea University
Prifysgol Abertawe
Medical School
Ysgol Feddygaeth

The Concept Library Usability Scale

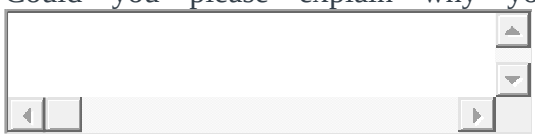
An example of a saved EHR Phenotype in the CALIBER research platform



* 1. I think that I would like to use this concept library frequently.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

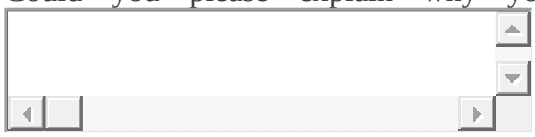
Could you please explain why you disagreed or agreed with this statement?



* 2. I think that I would need the support of a technical person to be able to use this concept library.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Could you please explain why you disagreed or agreed with this statement?



* 3. I found the various functions in this concept library, such as searching and viewing concepts; and creating and editing concepts, were easy to use.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 4. I felt very confident using the concept library.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 5. I needed to learn a lot of things before I could get going with this concept library.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 6. I think that the user documentation is task oriented and consists of clear, step by step instructions.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

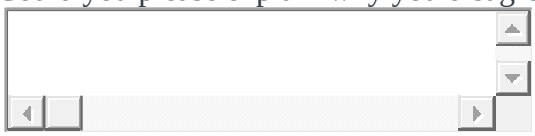
* 7. I found the concept library supports advanced functional tasks (e.g., it allows using of programming languages such as R, SQL, or Python).

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 8. I feel it is acceptable if I am required to reference the concept library when publishing papers.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Could you please explain why you disagreed or agreed referencing the concept library?



* 9. I found that it was easy to understand how the concept library is run and managed.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 10. I thought the concept library supports clear algorithms labelling convention.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 11. Can you tell us more about why you give the answers you did? (e.g., what could be improved? and what did you like about the system?)

12. If you are happy to participate in one-to-one interview, please provide us with your contact information.

Name
Company
Address
Address 2
City/Town
County
Post Code
Country
Email Address
Phone Number

Thank you for completing the Survey.

3.7.5 Ethical approval

I modified the previous Application Form for Standard Ethical Approval (See appendix 3), which was submitted for the purpose of conducting the one-to-one interview and focus group discussions, and I requested an approval to conduct the quantitative study (See appendix 5). I followed the ethical guidelines of the Swansea University Medical School's Research Ethics Sub-Committee (RESC), which include the following: 1) informed consent, 2) ensuring participants confidentiality and anonymity, and 3) ethical reporting of findings.

3.8 THE THIRD PHASE: CONDUCTING OF THE QUANTITATIVE STUDY USING THE CONCEPT LIBRARY USABILITY SCALE

3.8.1 Validation and Pilot test the Concept Library Usability Scale

To examine the content consistency and validity of the Concept Library Usability Scale, an e-mail survey instrument that I created, I contacted experts who were familiar with my research topic and had experience working with concept libraries and linked electronic data repositories, including a professor and an academic doctor, and asked them to read through the Concept Library Usability Scale and assess whether the statements effectively captured the topic under research. Accordingly, necessary changes have been made. The e-mail survey

was tested with some participants to ensure that the questions are understandable to the users of concept libraries, who are the target participants for the survey.

3.8.2 Procedure

The participants and their emails (N= 200) were identified by searching on https://cprd.com/bibliography for authors who published papers in 2020 and 2021 using CPRD data. I described the purpose of the research in the e-mail invitation to the participants, guaranteed them of their anonymity, and clarified to them that they were under no obligation to participate and could withdraw out at any time. After completing the surveys, the participants were asked to explain any ideas or thoughts elicited by the questions of the e-mail survey instrument and to make any recommendations to improve it. This qualitative component, that was used in combination with the quantitative component, allowed us to reassess the overall characteristics of the e-mail survey instrument.

3.8.3 Measures

Various rating scales have been created to directly measure attitudes or opinions, which means a survey participant is aware that their attitude is being examined. The Likert Scale (1932) is the most commonly used [141]. We utilised the Likert Scale because it allows the participants to express their level of agreement or disagreement by completing the Concept Library Usability Scale, an e-mail survey instrument we developed. It is a self-reported 12-statement scale that examines the usability of existing concept libraries for disease phenotyping. We used the five Likert Scale categories, which are: “Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, and “Strongly agree”, to assess the first 10 specific statements. We used the SurveyMonkey tools available at (https://www.surveymonkey.com) in order to analyse the findings. In the survey's design section, we used the Likert Scale, which is a variation of the matrix statement that allows us to assign weights to each answer choice. Below are the steps that we followed to incorporate this type of matrix statement into our email survey (the Concept Library Usability Scale):

1. From the builder section of the sidebar, we dragged and dropped the Matrix/Rating Scale into our survey.

2. We inserted the statements we wanted the participants to evaluate in the rows’ fields.

3. We entered the responses (i.e., “Strongly disagreed”, “Disagreed”, “Neither agreed nor disagreed”, “Agreed”, and “Strongly agreed”) that we wanted participants to use to evaluate the row statements in the columns’ fields.

For each response choice in the survey's analyse results section, the Likert Scale automatically generated a weighted average. When we inserted the statements, we selected the (Use Weights) option to see the average rating for the statements the participants were asked to evaluate. The following are the steps we took to enable the automatic calculation of the total score of the Concept Library Usability Scale (0 – 100):

1. We converted the scale to a numerical value for each of the first ten statements.

2. For each positive statement, we assigned a maximum point value of 10 to the strongly agreed response and a minimum point value of 2 to the strongly disagreed response as follows:

- Strongly Disagreed: 2 points
- Disagreed: 4 points
- Neither agreed nor disagreed: 6 points
- Agreed: 8 points
- Strongly Agreed: 10 points

3. In contrast, for each negative statement, we assigned a minimum point value of 2 to the strongly agreed response and a maximum point value of 10 to the strongly disagreed response as follows:

- Strongly Disagreed: 10 points
- Disagreed: 8 points
- Neither agreed nor disagreed: 6 points
- Agreed: 4 points
- Strongly Agreed: 2 points

To gain a thorough understanding of the participants' responses, we included comment boxes below the following statements: 1, 2, and 8 that allowed participants to describe why they agreed or disagreed with these specific statements. To encourage the participants to provide additional details, I have added a large comment box for item 11 so that they would have an opportunity to explain their overall responses to all statements of the survey. Barnum mentioned that the comments of the participants can provide valuable insights about their experiences, both positive and negative [137]. At the end of the survey (item 12), I requested them to provide us with their contact information if they were willing to participate in a one-on-one interview.

3.8.4 Sampling approach of the participants

Creswell and Clark stated that while samples for both the qualitative and quantitative phases should ideally come from the same population, the quantitative phase's sample size is often substantially bigger [99]. Quantitative research is commonly used to measure concerns and generalise study findings to the broader population, as a result, a large sample size and random selection of participants are necessary [120].

I sent e-mail invitations to the participants (N= 200) twice, which included researchers, health professionals, and clinicians asking them to complete a short e-mail survey. The purpose of this e-mail survey was to explore their attitudes and opinions about the usability of one of the existing concept libraries in the UK (i.e., the CALIBER research platform), which contains ‘research ready' variables derived from inked electronic health records (EHRs) from primary care, coded hospital records, social deprivation data, and cause-specific mortality data. I assured participants that all data obtained would be completely anonymous. (Figure 3.8 shows the e-mail invitations to the participants)

The participants were instructed in the e-mail invitations to spend approximately fifteen minutes searching for algorithms that they would like to use for their future research or that they had recently used for their studies using the CALIBER research platform link (https://www.caliberresearch.org/portal/phenotypes) and then provide feedback on the resource's usefulness in this case by completing our short e-mail survey, The Concept Library Usability Scale, using the following link (https://www.surveymonkey.com/r/7NDZPN9). The requirement for simplicity and speed in order to motivate participants to finish the survey so that there would be enough data to measure subjective reactions to a concept library usability.

Figure 3.6: The e-mail invitations to the participants

Dear (the name of the participant),

We wanted to learn about people's experiences with phenotype concept libraries (e.g., a list of READ codes/ICD 10 codes for identifying conditions), and we wanted to ask if you would be willing to do a short survey on using the CALIBER research platform.

CALIBER is a research platform, which is an example of existing concept libraries in England that contains 'research ready' variables collected from linked Electronic Health Records (EHRs) from primary care, coded hospital records, social deprivation data, and cause-specific mortality data. It is an open-access resource that provides information, tools, and phenotyping algorithms for UK electronic health records data, available through the CALIBER research platform, to the research community. We would like to know your views and thoughts on this resource. We would like to ask if you would spend about 15 minutes using it and giving us some feedback, and all the information collected from you will be anonymous. Please would you search for algorithms that you would like to use in the future or that you have used recently and give us feedback on how useful the resource was in this case by doing the short survey below.

- Link to the website of the CALIBER research platform: https://www.caliberresearch.org/portal
- Link to the EHRs Phenotypes: https://www.caliberresearch.org/portal/phenotypes
- Link to the short survey: https://www.surveymonkey.com/r/7NDZPN9

Thank you for your time and feedback (and if you know other people who might use concept libraries, please would you be willing to forward this email to them too)

With thanks and very good wishes

Best Regards,

Zahra Almowil

M.Sc. in Health informatics

Data Science Building

School of Medicine, Swansea University

Swansea, UK

Chapter 4

Review 1:

The purpose of this chapter is to review the literature on existing concept libraries for disease phenotyping, which serve as platforms for multiple researchers to store, manage, and share phenotypes (diagnoses, symptoms, medications, and procedures). This review aims to examine how they are used and identify current gaps and future development. This chapter is based on the following published paper in the International Journal of Population Data Science on 16/06/2021: Concept libraries for automatic electronic health record-based phenotyping: A review, which can be accessed via https://doi.org/10.23889/ijpds.v6i1.1362

Authorship Declaration	
Candidate	**Name and College**
Author 1	Zahra Almowil, Data Science Building, Medical School, Swansea University, Wales SA2 8PP
Author 2	Shang-Ming Zhou, Centre for Health Technology, Faculty of Health, University of Plymouth, Plymouth, PL4 8AA, UK
Author 3	Sinead Brophy, Data Science Building, Medical School, Swansea University, Wales SA2 8PP

Author Details and their Roles:

Paper 1 (Concept libraries for automatic electronic health record-based phenotyping: A review). It is located in Chapter 4.

The candidate conceived and designed the research, collected and analysed the data, and wrote the first draft and re-drafts of the paper. Her work was supervised by Shang-Ming Zhou and Sinead Brophy. We also independently read and agreed the data extract. SZ advised on the analysis. The work in this paper was 85% that off the candidate and 15% that of others.

We the undersigned agree with the above stated “proportion of work undertaken” for each of the above published peer-reviewed manuscripts contributing to this thesis:

Signed Candidate

Author 1: Zahra Almowil
Author 2: Shang-Ming Zhou
Author 3: Sinead Brophy

International Journal of Population Data Science

Journal Website: www.ijpds.org

Concept libraries for automatic electronic health record based phenotyping: A review

Almowil, Zahra A[1], Zhou, Shang-Ming[2], and Brophy, Sinead[1*]

Submission History

Submitted:	26/06/2020
Accepted:	02/10/2021
Published:	15/06/2020

[1]Swansea University Medical School, Wales SA2 8PP
[2]Centre for Health Technology, Faculty of Health, University of Plymouth, Plymouth, PL4 8AA, UK

Abstract

Introduction
Electronic health records (EHR) are linked together to examine disease history and to undertake research into the causes and outcomes of disease. However, the process of constructing algorithms for phenotyping (e.g., identifying disease characteristics) or health characteristics (e.g., smoker) is very time consuming and resource costly. In addition, results can vary greatly between researchers. Reusing or building on algorithms that others have created is a compelling solution to these problems. However, sharing algorithms is not a common practice and many published studies do not detail the clinical code lists used by the researchers in the disease/characteristic definition. To address these challenges, a number of centres across the world have developed health data portals which contain concept libraries (e.g., algorithms for defining concepts such as disease and characteristics) in order to facilitate disease phenotyping and health studies.

Objectives
This study aims to review the literature of existing concept libraries, examine their utilities, identify the current gaps, and suggest future developments.

Methods
The five-stage framework of Arksey and O'Malley was used for the literature search. This approach included defining the research questions, identifying relevant studies through literature review, selecting eligible studies, charting and extracting data, and summarising and reporting the findings.

Results
This review identified seven publicly accessible Electronic Health data concept libraries which were developed in different countries including UK, USA, and Canada. The concept libraries (n = 7) investigated were either general libraries that hold phenotypes of multiple specialties (n = 4) or specialized libraries that manage only certain specialities such as rare diseases (n = 3). There were some clear differences between the general libraries such as archiving data from different electronic sources, and using a range of different types of coding systems. However, they share some clear similarities such as enabling users to upload their own code lists, and allowing users to use/download the publicly accessible code. In addition, there were some differences between the specialized libraries such as difference in ability to search, and if it was possible to use different searching queries such as simple or complex searches. Conversely, there were some similarities between the specialized libraries such as enabling users to upload their own concepts into the libraries and to show where they were published, which facilitates assessing the validity of the concepts. All the specialized libraries aimed to encourage the reuse of research methods such as lists of clinical code and/or metadata.

Conclusion
The seven libraries identified have been developed independently and appear to replicate similar concepts but in different ways. Collaboration between similar libraries would greatly facilitate the use of these libraries for the user. The process of building code lists takes time and effort. Access to existing code lists increases consistency and accuracy of definitions across studies. Concept library developers should collaborate with each other to raise awareness of their existence and of their various functions, which could increase users' contributions to those libraries and promote their wide-ranging adoption.

Keywords
linked Electronic health records, phenotype algorithms, concept libraries, repeatable research, review

*Corresponding Author:
Email Address: S.Brophy@swansea.ac.uk (Sinead Brophy)

https://doi.org/10.23889/ijpds.v6i1.1362

Introduction

Electronic health records (EHR) have been adopted across the UK. For example, in terms of primary care in the UK there are the following four databases: 1) CPRD (Clinical Practice Research Data Link); 2) THIN (The Health Improvement Network); 3) QResearch and 4) SAIL (Secured Anonymised Information Linkage) in Wales [2]. In addition, secondary care data such as the hospital admission system (HES – England, PEDW – Wales, SMR -Scotland), are linked to primary care records [1 - 4]. Such linked information creates the opportunity to undertake research into the causes and outcomes and pathway of disease. However, using linked routine data requires some specialist skills, for example, using the data requires: 1) identifying conditions of interest from diagnosis, treatments, and procedures, and 2) creating phenotype algorithms (such as diagnosis of rheumatoid arthritis and medication for rheumatoid arthritis) and developing specific inclusion and exclusion criteria [5].

The construction of phenotype algorithms enables repeatable research and ensures that different researchers are using the same standards to identify patients [6]. However, the process of constructing phenotype algorithms is very time consuming and resource costly [7], and so reusing previously created phenotype algorithms to conduct repeatable research becomes a compelling solution. However, it is not common for researchers to share their clinical code lists in their published studies [8]. Therefore, it is difficult to make comparisons between studies as different studies often have different definitions of the same condition [4].

Although clinical code lists were published along with some EHR based studies, researchers often find it difficult to extract the relevant parts from lists for other research studies. Consequently, it is difficult to evaluate the transparency of EHR based research [9]. Even though researchers request better transparency in publishing clinical code lists [10, 11], currently journals and funding parties do not make it mandatory to publish code lists [9].

To address these challenges and ensure scientific transparency, data linkage centres have developed concept libraries for disease phenotyping, working as platforms to enable storing, managing, and sharing of phenotypes (Diagnoses, Symptoms, Medications and Procedures) by multiple researchers. For example, ClinicalCodes.org and CALIBER in the UK, and The Concept Dictionary and Glossary in Canada [9, 12, 13]. In the literature, concept libraries for disease phenotyping have different names and various definitions. We aim to review the literature of existing concept libraries to examine how they are used, identify the current gaps and future development. By evaluating the existing concept libraries and scoping what is missing in the current environment, this study could facilitate the development and improvement of concept libraries.

Methods

The five-stage framework of Arksey and O'Malley was used for the literature search [14]. This approach included defining the research questions, identifying relevant studies through literature review, selecting qualified studies, charting and collecting data, and summarising and reporting the findings.

Data Sources

This stage involved identifying the research questions, which provided the roadmap for subsequent stages. The questions to be addressed were:

- What concept libraries already exist?
- What are their features? Are there similarities or differences among them?

Identification of relevant studies

This stage involved identifying the relevant studies and developing a decision plan for where to search, which terms to use, which sources are to be searched, time span, and language. Searching was limited to peer reviewed manuscripts which were written in the English language and were published from 2010 to 2019. Five databases were searched including Medline, CINAHL, LISTA, Google Scholar, and Web of Science using the following sets of key words:

1. "electronic health record*" or "electronic medical record*" or "computerized health record*" or "computerized medical record*" or EHR or EMR
2. portal* or platform* or repositor* or library* or dictionary*
3. phenotyp* or e-phenotyp* or phenomic* OR "clinical code list*" or "clinical code*" or "clinical concept*" OR "clinical code set*" or "clinical value set*"
4. The sets of key words have been altered to be used in Google Scholar as recommended by this database as follows: ("electronic health record*" or "electronic medical record*" or EHR or EMR) AND (phenotyp*) AND (portal* or platform* or repository* for library* or dictionary*)

Selecting of eligible studies

The first author reviewed all the abstracts of the identified manuscripts (n=239) based on their relevance to the research questions. Those with relevant abstracts were taken forward to full assessment (n=50). Out of the fifty fully assessed manuscripts, only seven were selected as they matched the planned inclusion and exclusion criteria. The inclusion criteria for the selection process were to include manuscripts about public concept libraries for electronic linked health data-based phenotyping, and their different definitions, types, and functions, such as allowing users to share, reuse, and verify research methods (e.g., code lists, algorithms, and metadata). Manuscripts related to electronic health record phenotyping authoring tools are excluded. Figure. 1 depicts more information about the selection process of the related studies, and Table. 1 presents an overview of the seven concept libraries including their definitions/purposes, electronic data sources, coding systems, and examples of phenotype definitions in the seven public concept libraries.

Table 1: An overview of the seven concept libraries

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
1.General libraries					
ClinicalCodes.org	An online repository that contains a set of published studies. For each study a code list or a group of code lists has been uploaded on the ClinicalCodes.org site. Code lists are publicly accessible to improve validity and reproducibility of electronic medical record studies.	The University of Manchester. Institute of Population Health, UK	9. Spring ate DA, Ketopantoic E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. 2014; 9(6):6–11. https://clinicalcodes.rss.mhs.man.ac.uk/	Primary and secondary care using Read, OXMIS, SNOMED, CPRD, product/medical code, BNF code, ICD-9, ICD-10	Research article: Are symptoms of insomnia in primary care associated with subsequent onset of dementia? A matched retrospective case-control study, Link to the shared phenotypic descriptions at: https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/article/78/
CALIBER data portal	An open online repository of phenotyping algorithms that contains all definitions of research variables using CALIBER data sources in order to encourage research and promote transparency.	Led from the University College London (UCL) Institute of Health Informatics, UK	21. Dewaxes S, Gonzalez-Inquired A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019; 26(12):1545–59. https://www.caliberresearch.org/portal/phenotypes	Primary care, hospital records, social deprivation information, cause-specific mortality data. Using Read codes (a subset of SNOMED-CT), ICD-9, ICD-10, OPCS-4 (analogous to Current Procedural Terminology terms) and Gemscript.	Abdominal Hernia: "At the specified date, a patient is defined as having had Abdominal Hernia If they meet the criteria for any of the following on or before the specified date. The earliest date on which the individual meets any of the following criteria on or before the specified date is defined as the first event date: Primary care 1. Abdominal Hernia diagnosis or history of diagnosis or procedure during a consultation OR Secondary care 1. ALL diagnoses of Abdominal Hernia or history of diagnosis during a hospitalization

3

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
					OR Secondary care (OPCS4) 1. ALL procedures for Abdominal Hernia during a hospitalization" Link to the shared phenotypic descriptions at: https://www.caliberresearch.org/portal/phenotypes/chronological-map
The MCHP Concept Dictionary and Glossary	The Concept Dictionary includes comprehensive operational definitions and programming code for measurements used in MCHP research including a description of the problem(s) involved, methods used, and programming tips/ cautions, and the Glossary records terms that are widely used in research based on population. The Concept Dictionary was developed to help researchers use reliable, validated algorithms to perform methodologically comprehensive research.	The Manitoba Centre for Health Policy (MCHP), Canada	13. Ostapyk T. Manitoba Centre for Health Policy Data Repository. In: Michalos AC (eds) Encyclopedia of quality of life and well-being research. Springer, Dordrecht; 2014. http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/concept_dictionary.html	- The MCHP databases: Health, Education, Social, Justice, Registries, Support Files. - Operational definitions and SAS program code for variables or measures developed from administrative data. - The International Classification of Disease (ICD) diagnoses or ICD / CCI (Canadian Classification of Health Interventions) procedure / intervention Link to the shared	Manitoba Asthma Algorithms The following is an example of asthma algorithm developed by a research project. "Raymond et al. (2011) use a broader scope in their definition for asthma, defining it as one physician claim OR one hospital claim with a corresponding diagnosis of: ICD-9-CM: 464, 466, 490, 491, 493 or ICD-10-CA: J04, J05, J20, J21, J40, J41, J42, J45, J441, J448 OR one prescription for an asthma medication in a three-year period."

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
					phenotypic descriptions at: http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1305#a_references
Phenotype knowledgebase (PheKB)	An online environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms. The PheKB was designed to facilitate the transportability of algorithms into various research applications across different organizations, health care systems, and repositories of clinical data.	Led by Vanderbilt University, (the eMERGE Network Coordinating Centre), USA	18. Kirby JC, Speltz P, Rasmussen L V., Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalogue and workflow for creating electronic phenotype algorithms for transportability. J Am Med Informatics Assoc. 2016; 23(6):1046–52. https://phekb.org/	Clinical and genomic data from electronic health records. HCPT Codes, ICD 10 Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	Urinary Incontinence The cohort is defined with the following criteria: a. EHR of all male patients of 35 years of age or more, AND b. For which there is an ICD-9-CM / ICD-10-CM diagnosis of prostate cancer, AND c. For which there are at least two encounters before first treatment, AND d. For which there is at least one clinical note before first treatment, AND e. For which there is either prostatectomy surgery or radiation procedure performed as identified by CPT codes. Link to the shared phenotypic descriptions at: https://phekb.org/phenotype/1404
2. Specialized libraries					
Genome-Phenome Analysis Platform (GPAP)	An online data platform, where data from sequencing experiments contributed by collaborating research projects is processed using a standard pipeline and made accessible to registered users for online analysis through a user-friendly interface.	It was developed by RD-Connect and Led by Aix-Marseille University Medical School (AMU), France	25. Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-Connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014; 29(SUPPL. 3):780–7.	Genomic and clinical data from RD-Connect's partners rare disease-based research projects. The PhenoTips database stores phenotypic profiles for individual cases coded by human phenotype ontology (HPO).	Case 1: description RD-Connect identifier: Case1C Gender: Male, Age: 5 years, Referral: Congenital myasthenic syndrome,

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
			https://dx.doi.org/10.1007%2Fs11606-014-2908-8	A directory of biobanks and patient registries and a bio sample catalogue.	Onset: Congenital, Global pace of progression: Progressive (slow), Main clinical features: Neonatal hypotonia, Distal arthrogryposis, Inability to walk, Recurrent lower respiratory tract infections. Link to the shared phenotypic descriptions at: https://playground.rd-connect.eu/
The PhenoScanner V2	A database that contains publicly existing results of large-scale genomic association studies. It was developed to facilitate the cross-referencing of genetic variants with a wide variety of phenotypes for better comprehension of biology and pathways of disease.	The Cardiovascular Epidemiology Unit, University of Cambridge, UK	22. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019; 35(22):4851–3. http://www.phenoscanner.medschl.cam.ac.uk/	137 genotype–phenotype association datasets, including results for anthropometric traits, blood pressure, lipids, cardiometabolic diseases, renal function measures, glycemic traits, inflammatory diseases, psychiatric diseases and smoking phenotypes. It also includes the NHGRI-EBI GWAS catalogue, and dbGaP catalogues of associations.	Trait: Crohn's disease "A gastrointestinal disorder characterized by chronic inflammation involving all layers of the intestinal wall, noncaseating granulomas affecting the intestinal wall and regional lymph nodes, and transmural fibrosis. Crohn disease most commonly involves the terminal ileum; the colon is the second most common site of involvement.

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
					A chronic transmural inflammation that may involve any part of the DIGESTIVE TRACT from MOUTH to ANUS, mostly found in the ILEUM, the CECUM, and the COLON. In Crohn disease, the inflammation, extending through the intestinal wall from the MUCOSA to the serosa, is characteristically asymmetric and segmental. Epithelioid GRANULOMAS may be seen in some patients." Link to the shared phenotypic descriptions at: https://www.ebi.ac.uk/gwas/efotraits/EFO_0000384
Genotypes and Phenotypes Database (dbGap)	A National Institute of Health-sponsored repository tasked with archiving, curating and distributing information provided by studies examining genotype and phenotype interactions. It was developed with standardized identifiers that allows published studies to address or cite the primary data in a clear and uniform way.	National Centre for Biotechnology Information, USA	15. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's database of genotypes and phenotypes: DbGaP. Nucleic Acids Res. 2014; 42(D1):975–9. https://www.ncbi.nlm.nih.gov/gap/	Genetic and phenotypic databases sponsored by NIH and other agencies around the world including: Genotype, phenotype, exposure, expression array, epigenomic and pedigree data from genome-wide association studies (GWAS), sequencing studies and other large-scale genomic studies.	"Autism_Genome_Project_Subject_Phenotypes: The subject phenotype table includes data collected on sociodemography (n=2 variables; sex and European ancestry) and psychological and psychiatric observations (n=8 variables; spectrum and strict definition of autism, whether the subject is non-verbal and/or verbal, has low or high IQ, and the age of their first word and phrase).

7

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
					This table now also includes the stage of the study in which the individual was present and the whether individual is a member of a multiplex or simplex family." Link to the shared phenotypic descriptions at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/variable.cgi?study_id=phs000267.v5.p2&phv=161303&phd=3659&pha=3690&pht=2305&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1

Extraction, charting, and synthesis of data

Data was extracted from the seven related manuscripts using a data-charting form. A narrative review method was used to extract data about the investigated seven public concept libraries for electronic linked health data-based phenotyping including their names, types, and characteristics such as enabling users to share, validate, and reuse of research methods such as algorithms.

Collecting, summarising and reporting the findings

A thematic construction was used to provide an overview of the breadth of the literature, and then a thematic analysis was used to generate the results. The different types and the characteristics of the seven public concept libraries were summarised. The types of electronic data sources used in each library (e.g., primary or secondary care or genetic data) and the used coding system (e.g., Read, OXMIS, ICD-9, and ICD-10) were all reported.

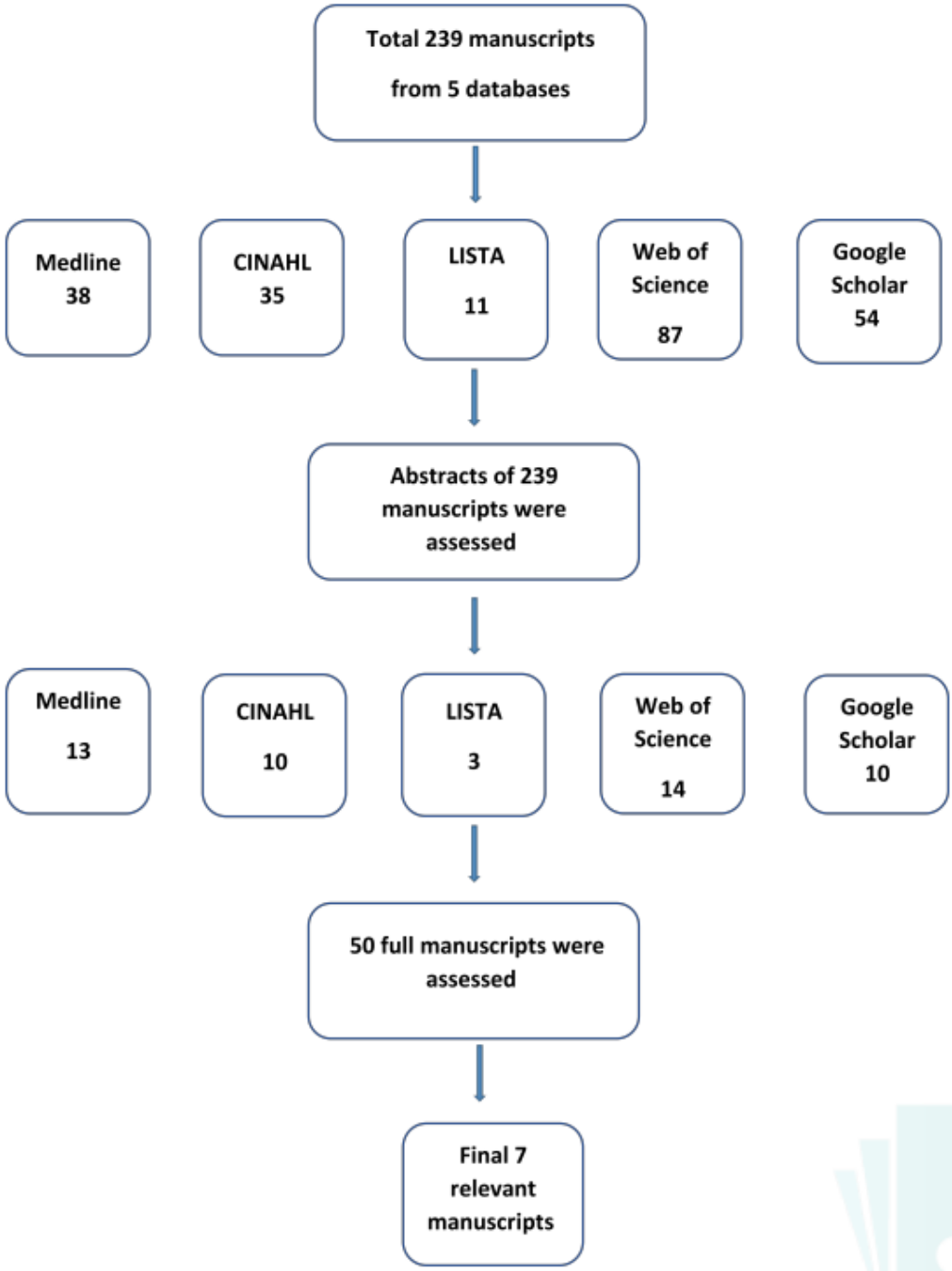
Results

Identified public concept libraries from the literature

There were seven public concept libraries from the literature developed by different countries including UK, USA, and Canada. These libraries were the ClinicalCodes.org [9], the Genotypes and Phenotypes Database (dbGaP) [15], Phenotype knowledgebase (PheKB) [18], the Manitoba Centre for Health Policy (MCHP) Concept Dictionary and Glossary [20], Clinical Disease Research using Linked Bespoke Studies and Electronic Health Records (CALIBER) [21], the PhenoScanner V2[22], and The Genome-Phenome Analysis Platform (GPAP) [24]. Four of the libraries were general libraries and held concepts and phenotypes on multiple specialities ranging from specific conditions (such as codes to identify lupus) to general demographic concepts (such as smoking status). Three of the libraries were specialised libraries and only give concepts on certain defined areas such as rare diseases. However, in common across all the libraries was that they allowed users to share by uploading their own concepts, to examine validity of concepts by showing where they were published, and all had the aim of facilitating reuse of research methods such as clinical code lists or metadata.

There were some clear differences between the general libraries such as archiving data from different electronic sources (e. g. primary care, secondary, social deprivation information, cause-specific mortality data, health, education, justice, and registries); using various types of coding systems (e.g. SNOMED, BNF, READ, ICD9 /10, and CCI (Canadian Classification of Health Interventions) [9] [18] [20] [21]; having different policies that govern accessing the underlying data sources (e.g. a researcher has to complete the Data Access Process (DAP) of the Manitoba Centre for Health Policy (MCHP) to access the data and conduct research by using the Manitoba Population Research Data Repository) [20]; and allowing different searching queries such as simple or more advanced searches (e.g. CALIBERcodelists package enables

Figure 1: Overview of the steps taken in the priority-setting process.



users to search for code lists by synonym or code stub, and combine search terms using Boolean operators) [29].

However, they share some clear similarities such as having similar purposes (e.g. helping researchers to perform comprehensive research, promoting transparency in sharing research methods, and improving reproducibility of studies; enabling users to upload their own code lists and other important documents (e.g. users of PheKB can upload related documents and their phenotypes along with multidimensional metadata labels, documents including detailed descriptions of the computable algorithms such as types of used data, logic of execution, definitions of data, and flow charts) [18]; and allowing users to use/download the publicly accessible code lists (e.g. users of ClinicalCodes.org can download a file containing all codes associated with a study as csv files) [9].

There were some differences between the specialized libraries such as using data that were generated by various electronic databases (e.g. the PhenoTips database, database of biobank and patient registries, the NHGRI-EBI GWAS catalogue, and genetic and phenotypic databases sponsored by NIH and other agencies around the world) [15] [22] [25]; allowing diverse searching strategies (e.g. registered users of GPAP may select one or more individuals such as trios or other family relationships to explore and then filter and refine the outcomes by inheritance mode, population frequencies, tools for in silico pathogenicity prediction, gene lists and Linvar, HPO and OMIM codes)[30]; and enabling different searching queries such as simple or complex searches (e.g. all publicly released dbGaP studies can be queried by users. Queries can be very simple, just a keyword of interest ('cancer') or complex, making use of search fields and Boolean operators ('cholesterol[variable] AND phs000001') [15].

Conversely, there were some similarities between the specialised libraries such as enabling users to share by uploading their own concepts in the libraries to analyse the validity of concepts by showing where they were published (e.g. the GPAP enables clinicians and researchers who upload patient datasets to analyse their own data [30], and NIH-funded researchers can share their produced data, in the dB Gap) [15]; all aimed to encourage the reuse of research methods such as lists of clinical code or metadata (e.g. registered users of GPAP are allowed to access and search data sets provided by other researchers on similar patients[30], and users of the PhenoScanner V2 can use the archived findings from large-scale genetic association studies which are publicly accessible) [22]; and allowing access to some datasets through specific established control access (e.g. individual level data is accessible in the dbGap to scientists around the world through controlled application of access) [15]. Information about all the seven concept libraries such as their access URL and references of the seven manuscripts are presented in Table 1.

An overview of some the seven public concept libraries' features

1. **Names and Definitions:**

   Each of the investigated concept libraries has a specific name and a unique definition (Table 2). For example, CALIBER is defined as "a unique research platform consisting of 'research ready' variables extracted from linked electronic health records (EHR) from primary care, coded hospital records, social deprivation information and cause-specific mortality data in England" (https://www.ucl.ac.uk/health-informatics/caliber). Whereas, the Database of Genotypes and Phenotypes (dbGap) is defined as "a National Institutes of Health-sponsored repository charged to archive, curate and distribute information produced by studies investigating the interaction of genotype and phenotype" [15].

2. **Types**

   2.1 The general concept libraries:

   The ClinicalCodes online repository

   The ClinicalCodes repository contains a selection of published studies that have been uploaded to the ClinicalCodes.org site along with a code list or a series of code lists. A code name, coding system (Read, OXMIS, SNOMED, CPRD product / medical code, BNF code, ICD-9, ICD-10), definition and type of entity (diagnostic, drug, examination, clinical sign, administrative, demographic, observational, immunization) are assigned to all individual clinical codes. Metadata and links to studies code lists are accessible as research objects that could be shared in machine-readable form throughout platforms. A research object file of JavaScript Object Notation (JSON) is available for each study that contains metadata (title, author, abstract, reference, link, DOI), commentary on the study level, commentary on the code list level and links to the individual files of the code list. Such object research files are directly accessible when inserting a '/ro' to the URI for a study e.g., (www.clinicalcodes.org/medcodes/article/5/ro) [9]. The developers of the ClinicalCodes repository have created an open-source R package (rClinicalCodes) to automate the downloading and importing lists of clinical code and metadata through the research object file from the repository website: (https://cran.r-project.org/web/packages/rClinicalCodes/index.html. The developers of the ClinicalCodes repository will implement in the future: 1) Searching and downloading of codes by disease group, keyword and/or code group. 2) Methods for downloading code-lists and article metadata in machine readable form. 3) An API for downloading code-lists programmatically [9].

   The Clinical Research using Linked Bespoke Studies and Electronic Health Records (CALIBER)

   CALIBER has developed the CALIBERcodelists package to manage ICD-10, Read and OPCS coding lists to identify medical conditions for research using CALIBER or other UK electronic health record databases. The package is written in R language, but many of the functions are accessible through an interactive menu and do not require any experience with R. The package has many features: 1) provides a standardized approach to identify codes of interest including Read, ICD-10 and OPCS through searching for term text or codes, 2) enables displaying of code lists on a spreadsheet and removing individual terms or modifying their categories, 3) downloads

code lists in a variety of formats, and uploads them in default file format, 4) allows comparing of one code list to another, or combine two code lists together, 5) enables converting of code lists across dictionaries using the NHS mapping between the terminologies of Read/OPCS and Read/ICD-10, 6) processes a document which contains code to produce a code list and a descriptive text, and produces a comprehensive HTML document and a standardised format code list [27].

The Manitoba Centre for Health Policy (MCHP) Concept Dictionary and Glossary

MCHP has built a range of web-based tools that record the historical usage of the repository-saved information such as the MCHP Concept Dictionary and Glossary [20]. Although there are short definitions for widely used terminology in the glossary, the Concept Dictionary includes comprehensive operational definitions and programming code for measurements used in MCHP research. A coherent documentation approach is used to describe the research methodologies. They can be presented either as best practices, or as different versions of historical overviews. Enhancements substitute older versions with a best-practice approach which requires authoritative approval of what is "the best", and the historical overview records all published methods for making options accessible to the user and information about the methodologies used in previous studies [26].

The Phenotype Knowledgebase (PheKB)

The Phenotype Knowledgebase (PheKB accessible at http://phekb.org) was created within the eMERGE Network as a workflow management system and learning centre to support computable algorithm creation, validation, and sharing. It enables the transportability of algorithms across various research applications, multiple organizations, health care systems, and clinical data repositories through feedback processes and standardised implementation performance measures. PheKB contains built-in tools designed specifically to improve sharing of knowledge across sites, for example, the Data Dictionary / Data Validation Tool and the data management function. The Data Dictionary / Data Validation Tool is a registered user embedded resource that validates definitions of covariate data and related data, promotes data standardization, and early-stage quality assurance to exchange data for study sets efficiently. It is used for identifying errors and warnings in data dictionaries and data files related to a given phenotype through a custom Drupal module. It can show errors and warnings regarding the structure and content of the files as files are uploaded, while the data management function provides tracking tools for users to easily determine what data has been shared, what algorithm it is linked to, and by whom it was shared [18].

2.2 The specialized concept libraries:

The Genome-phenome analysis platform (GPAP)

The RD-Connect built an integrated Genome-phenome analysis platform (GPAP), which is a user-friendly tool for diagnosing and discovering genes [30]. It links anonymised omics and clinical data with tools and services to examine these data online. The main portal provides links to the genomics analysis interface and the Phenotypes database that store ontology of phenotypic profiles coded for individual cases by human phenotype (HPO). GPAP also includes a database of biobanks and patient registries, and a catalo of bio samples that allows information of individual samples housed in participating biobanks to be drilled down [17]. For example, a researcher may select one or more individuals (e.g., trios or other family relationships) to explore and then filter and refine the outcomes by inheritance mode, population frequencies, tools for in silico pathogenicity prediction, gene lists and Linvar, HPO and OMIM codes [25] [30].

The Pentosane V2

The developers of PhenoScanner V1 have collected more than 5,000 genotype-phenotype association datasets to create version 2 of the catalogue (PhenoScanner V2). PhenoScanner V2 has an API that contains an R package and a Python command line tool associated with it, which enables users to search for PhenoScanner V2 genotype-phenotype associations within R or from a terminal. All results, irrespective of P-value, can be presented when querying genetic variants, allowing the user to find indication against phenotype associations. PhenoScanner V2 has new features to facilitate improved 'phenome scans' including: 1) an expanded database of human genotype-phenotype associations divided into phenotype classes (diseases and traits, gene expression, proteins, metabolites and epigenetics), 2) new search selections such as gene, genomic region and queries based on phenotypes 3) linkage disequilibrium (LD) information for the five super-ancestries in 1000 Genomes 4) variant annotation and trait ontology mappings 4) annotation variations and ontology mappings of traits, and 5) a new Platform and API [22].

The database of Genotypes and Phenotypes (dbGaP)

The Genotypes and Phenotypes database (dbGaP) enables licensed users to identify and display different regions of the human genome, such as all the Allele frequencies and subgroups of individual-level genotype as well as sequence data, which are stored in that region in dbGaP, without accessing data sets of interest and performing multiple analyses [23]. The browser uses the standard graphical interface built for data from the 1000 Genomes Project and dbGaP by the National Centre for Biotechnology Information (NCBI), which incorporates sequence viewer track views with genotype tables and a novel sample / subject data selector showing core sample phenotype data [15]. The webpage of the browser includes a selection of 'widgets' pages showing data from the dbGaP view-only data project, which is data from the collection of dbGaP general research usage (GRU). The widgets work in such a way that one widget operation causes updating of other widgets on the page. See online browser documentation (https://www.ncbi.nlm.nih.gov/gap/ddb/help/) and The NCBI YouTube channel (https://www.youtube.com/user/NCBINLM) for additional widget information.

3. Characteristics

Some of the characteristics of the seven concept libraries are presented in Table 2.

3.1 Sharing of Concepts

All of the identified concept libraries allow researchers to share some or all of their research methods such as clinical codes list, metadata, and algorithms. For example, users of the ClinicalCodes.Org are able to upload the code list and metadata for specific codes, and add comments at the code list, or study level. However, an account should be created first [9]. According to the developers of ClinicalCodes.Org, "To date: 93375 clinical codes have been deposited over 521 code lists" [19]. Similarly, Phenotype knowledgebase (PheKB) enables researchers to upload related documents and their phenotypes along with multidimensional metadata labels including the methods used in the phenotype standards such as International Classification of Disease (ICD) codes, medications, and natural language processing (NLP). Researchers also can upload documents that include detailed descriptions of the computable algorithms, such as types of data used, logic of execution, definitions of data, and flow charts [18].

Builders of the Concept Dictionary and Glossary at the Manitoba Centre for Health Policy (MCHP) encourage researchers to share their discoveries, such as creating new concepts or updating existing concepts to grow and improve the value of this publicly accessible resource. Their Concept Dictionary describes more than 300 research concepts developed at MCHP for the analysis of data contained in the data warehouse hosted at MCHP [20]. Also, the developers of the CALIBER platform promote collaborative research. They compiled more than 90,000 terms from five standardised clinical terminologies to construct 51 validated phenotyping algorithms (35 diseases or syndromes, 10 biomarkers, 6 risk factors for lifestyles) [21]. All data sources are made accessible to researchers and can be accessed in a secure data-safe haven environment located at UCL IHI / Farr Institute, London or could be accessed remotely. Due to the varied clinical backgrounds of the datasets, they offer training on data sources, coding, consistency and management with the CALIBER team [16].

The Phenoscanner V2 database includes more than 5000 genetic association datasets from publicly accessible datasets of complete summary of associations findings collected by the NHGRI-EBI (https://www.ebi.ac.uk/gwas/downloads/summary-statistics) and NHLBI (https://grasp.nhlbi.nih.gov/FullResults.aspx), and recent literature reviews and GWAS omics datasets [22]. Also, the Genotypes and Phenotypes Database (dbGaP) allows sharing of information obtained from studies examining genotype and phenotype interactions [15]. These studies include research of genome, medical sequencing, molecular diagnostic assays, and correlation between genotype and non-clinical traits [23]. Similarly, the Genome-Phenome Analysis Platform (GPAP) facilitates data sharing as it now opens for submissions of projects from all users, and not only from RD-Connect partners [24]. One of their main objectives is to help the contributed projects to quickly make their data available to the broader community of rare disease researchers [25].

3.2 Validation of Concepts

Most of the identified concept libraries from the literature have described their validation methods either in their published studies, or in their websites, or in both of them. For example, in the CALIBER platform, EHR-derived phenotypes have been extensively validated using six different approaches: cross-EHR source concordance, case note review, consistency of risk factor-disease association from non-EHR studies, consistency with prior prognosis research, consistency of genetic association, and external populations. The builder of the platform acknowledged that the case study would inform which validation(s) are most important. For example, phenotyping algorithms developed for disease epidemiology (e.g., screening or disease surveillance) might be designed for higher sensitivity whereas those used in genetic association studies might be designed to maximize positive predictive value (PPV) [21].

The developers of the PheKB have developed the Data Dictionary / Data Validation Tool, which validates covariate data descriptions and related data and is a tool embedded for registered users. The user uploads to the phenotype-related page, and the tool verifies the data dictionary file for compliance with standards and best practises. A specified set of rules defines differences from the standard or guidelines for best practises [18].

The Concept Dictionary and Glossary was built at MCHP to assist researchers to carry out methodologically comprehensive research using consistent, validated algorithms [20]. According to their builders, concepts are written using original ideas and methods developed for MCHP reports, then reviewed and shaped according to common standards by the repository analyst [26]. Similarly, the data submitted including individual genomic and phenotype data, analytical results, general study information, are subject to quality checks by Genotypes and Phenotypes Database (dbGaP) staff before the Genotypes and Phenotypes Database (dbGaP) information is released publicly [15].

3.3 Reusing of Concepts

All of the seven concept libraries allow reusing of stored data, clinical code lists and algorithms, however each concept library has established certain terms and searching features for users. For example, any user can download code lists from the ClinicalCodes.org repository. Once deposited, code lists will be freely available, with no login needed to download the codes. In addition, an open-source R package has been developed to automate the downloading of code lists from the online repository [9]. Also, the CALIBER platform allows reuse of existing lists of codes by researchers. Users can access phenotyping algorithms defining over 90 diseases and metadata. CALIBER has CALIBERcodelists package [27], which enables users to search for code lists by synonym or code stub, allows users to combine search

terms using Boolean operators, and supports regular expressions for more advanced search queries. In addition, it allows downloading of the list of codes and some basic metadata for example, the name and version of the code list as a csv file [21].

Algorithms and multiple implementation results can be publicly viewed in the PheKB website when authors designate it as "final". By using metadata, users can search an algorithm based on inclusion or exclusion of data elements classes, such as diagnosis, author, or keyword. Currently, there are 414 users of PheKB from 52 different institutions. The median of used algorithms per institution is four. As of March 2020, PheKB include 30 public algorithms with 66 executions and 62 non-final algorithms with 83 executions in different stages of development [18].

PhenoScanner V2 is a searchable library of findings from large-scale genetic association studies which are publicly accessible. The database now includes more than 350 million association results and more than 10 million original results genetic variations [22]. The developers of the PhenoScanner V2 specified the terms of use in their website: 1) users should cite both their papers in any publication or presentation 2) users should cite the original paper where the results were obtained, including the references for the linkage disequilibrium statistics and variant & phenotype mappings where used and 3) users should comply with any other terms relating to the data [28].

The Concept Dictionary and Glossary at MCHP describes more than 300 research concepts developed at MCHP for the analysis of data contained in the data warehouse hosted at MCHP [29]. Over time, traffic on the MCHP website has increased. Their analysis software, Deep Log Analyser, recorded more than two million visits in 2018. In addition, the Charlson Comorbidity Index (CCI) including a glossary of terms and concepts has been ranked as the most widely viewed definition in the MCHP Concept Dictionary for many years. Similarly, the Elixhauser Comorbidity Index concept has regularly appeared among the top five most viewed concepts; whereas measures of comorbidity have often been among the most viewed concepts [26].

The Genotypes and Phenotypes Database (dbGaP) provides free access to publicly available information on completed research and studies-related documents. However, individual level data is open to scientists around the globe via controlled access application. This platform allows researchers and clinicians to quickly interpret and compare DNA sequencing data with clinical knowledge, including those who do not have training in bioinformatics. Information in the Genotypes and Phenotypes Database (dbGaP) is organized as a hierarchical structure and includes the accessioned objects, phenotypes (as variables and datasets), various molecular assay data, analyses and documents. The Genotypes and Phenotypes Database (dbGaP) enables both simple as well as advance searches [15].

4. **Limitations**

Some of the developers of the concept libraries mentioned some of their limitations as described below:

The ClinicalCodes repository does not offer methods for downloading code-lists and article metadata in machine readable form according to their developers, and is lacking search features needed to facilitate queries such as searching and downloading of codes by disease group, keyword and or code group, all of which are planned to be added in the future. Also, they stated that it does not have a protocol for enforcing quoting of the downloaded code lists. Therefore, it would be difficult to connect code lists from earlier studies [9].

The developers of CALIBER mentioned that there are some fairly complete measures in CALIBER's data, for instance, 82.6 percent of people with at least one measurement of BMI using the concepts in the library. But some measurements are less comprehensive, for example, only 44.9 percent have at least one total measure of cholesterol when using the library concepts. They also mentioned that different records in CALIBER can represent the same event or subsequent events at similar points in time. For example, fatal myocardial infarction can be reported in up to four diverse sources that vary in their specificity in diagnosis and precision in timing [21].

The developers of PheKB stated that some algorithms cannot work at a given site as well as at another, and validation is the only way to distinguish poorly performing algorithms. They also mentioned that PheKB does not have programmatic interfaces with some of the networks needed to enable fast exchange of executable phenotyping algorithms [18].

The developers of dbGaP declared that the National Institutes of Health (NIH) policies, such as restricting the context of available data only to researchers who consented to general research use, leads to limiting the related phenotype data to specific demographic and disease status information, reducing the data download capabilities of the browser, and showing a browser watermark that images must not be recorded or published [15] [31].

Discussion

Statement of main findings

Globally, the development and use of concept libraries is important for reusable health studies. A number of data linkage centres around the world have developed different concept libraries to facilitate repeatable research. This paper examined seven concept libraries, and variations in their definitions, names, types, functions, coding systems, and data access restrictions. One of our findings is that these concept libraries have developed independently and so are duplicating work but in slightly different ways.

For wide use of concept libraries, collaboration across data linkage centres is needed to develop common standards that govern and guide these emerging libraries. For example, they

Table 2: Some of the characteristics of the seven concepts libraries

Concept Libraries	Access to the underlying data sources	Sharing/Uploading of Concepts	Reusing/Downloading of Concepts
1.General libraries			
ClinicalCodes.org	In the top menu tab 'Browse published studies," the user may choose a published study from the list. It then shows all the code lists associated with that study.	"Users must register with ClinicalCodes.org (in the menu bar login/signup) and choose 'upload codes.' First, they need to add some metadata of the published study and then they can upload several codes lists as delimited text files into that study. Metadata and links to studies code lists could be shared in a machine-readable form using the available open-source R package (rClinicalCodes).	Code lists are released on ClinicalCodes.org using a Creative Commons Attribution 3.0 Unported License (CC BY 3.0), and a file containing all codes associated with a study can be downloaded and used freely by any user. Downloading individual code lists is a single-click process that does not involve logging in or supplying user information. Users can choose to explore and download some or all of the code lists as csv files.
CALIBER data portal	Access to CPRD linked data on the CALIBER portal complies with the governance policies for data access by the CPRD. Researchers should first sign agreements with UCL to access CPRD data. Non-UCL partners must apply for CPRD to become a CPRD-approved partner and sign a UCL-approved sub-license agreement.	If a project proposal for a researcher has been accepted, a registration with the UCL Identifiable Data Handling Service (IDHS) will be arranged in order to create a new share for the project on the safe haven. The data facilitator at CALIBER will direct researchers through the entire process.	All definitions of research variables that use CALIBER data sources are publicly available and can be accessed in human and machine-readable formats. CALIBER codelists package enable users to search for code lists by synonym or code stub, combine search terms using Boolean operators, and download the list of codes and some basic metadata as a csv file.
The MCHP Concept Dictionary and Glossary	The Data Access Process (DAP) of the Manitoba Centre for Health Policy (MCHP) are the processes that a researcher has to complete to access the data and conduct research using the Manitoba Population Research Data Repository.	Researchers can share their work, such as creating new concepts or updating existing concepts. The concept development guidelines are defined in the Concept Development Template. Concept Development Template.	**Concept Dictionary:** More than 300 research concepts developed at MCHP are publicly accessible. **Glossary:** Terms of documentations widely used in population-based research are freely available. Browse/Search the Concept Dictionary and Glossary
Phenotype Knowledgebase (PheKB)	Private Phenotypes with "In Development" status, phases of "Testing," or "Validation" are not publicly accessible, which can only be accessed if the user is logged in and the phenotype was shared with the user via one of the two collaborative groups: Owner Group Phenotypes or View Group Phenotypes.	Researchers can upload: Related documents and their phenotypes along with multidimensional metadata labels. Documents including detail descriptions of the computable algorithms, such as types of used data, logic of execution, definitions of data, and flow charts.	Algorithms and multiple implementation results can be publicly viewed in the PheKB website when author designated it as "final". By using metadata, users can search an algorithm based on inclusion or exclusion of data elements classes, such as diagnosis, author, or keyword.
2. Specialized libraries			
Genome-Phenome Analysis Platform (GPAP)	Only approved users who have completed the registration and verification process can access the data stored on the GPAP. Users must be affiliated with a recognized academic institution as accredited clinicians/researchers and must demonstrate their approval of the RD-Connect Code of Conduct by signing the Adherence Agreement.	Data sharing is open for project submissions from all users, not only from partners of RD-Connect, but they have to register first in the GPAP website. The GPAP enables clinicians and researchers who upload patient datasets to analyse their own data.	Registered users are allowed to access and search data sets provided by other researchers on similar patients. Registered users can match make, find second families, and find patient populations for validation studies.
The PhenoScanner V2	Some of the datasets are available for download including: dbSNP 147 with variant annotation from VEP, Linkage disequilibrium statistics from 1000 Genomes and a subset of the processed GWAS datasets, but users should first contact phenoscanner@gmail.com to get an approval.	Users can input one genetic variant, gene, genomic region or trait in the home page text box (www.phenoscanne.medschl.cam.ac.uk) or upload as a tab-delimited text file up to 100 genetic variants, 10 genes or 10 genomic regions.	"Users can use the archived findings from large-scale genetic association studies which are publicly accessible. Information provided by project members are unrestrictedly accessible.
Genotypes and Phenotypes Database (dbGaP)	Free access to information on completed studies are open to the public. Individual level data is accessible to scientists around the world through controlled application of access.	NIH-funded researchers can share their produced data. However, studies that are not sponsored by the NIH, individual NIH Institutes and Centres (IC) make judgments about whether non-NIH sponsored data should be accepted.	Open-access data can be accessed online or downloaded without prior authorization or permission from dbGaP. Individual level data download requests are handled through the dbGaP Authorized Access System (dbGaPAA), a web portal that manages request submissions, and enables safe high-speed large data download for authorized users.

should agree on a relatively standard definition/name for concept libraries to enable users to locate them and then use them easily. In addition, builders of concept libraries should cooperate with each other to increase awareness about their existence and their various functions. What one concept library might do that others do not do (e.g., provide SNOMED or BNF code lists or provide definitions for demographic variables such as smoking, BMI algorithms that other concept libraries do not do). Raising awareness of the features in the different libraries could increase the contributions of users to these libraries and accelerate their wide adoptions.

For a comprehensive adoption of concept libraries, their various functions, such as enabling users to share, validate, and reuse concepts (e.g., code lists), and their search features should be assessed by developers, funders, users, and experts to ensure that they meet the needs of various users including researchers, clinicians and data analysts. Since there are two different types of concept libraries 1) general libraries that hold phenotypes of multiple specialties 2) specialised libraries that manage only certain specificity such as rare diseases, users' preferences for the type of concept library types needs to be evaluated (e.g., through interviews, focus group, and surveys) before developing new concept libraries.

Strengths and limitations

This is the first study, to our knowledge, aimed at identifying existing concept libraries, exploring their various characteristics, and examining the current practises in this evolving field. Finding studies about a concept library for electronic health data phenotypes in the literature was challenging as there are a limited number of related studies. Another challenge was the lack of a standard name or definition that describes this kind of library. Therefore, a range of keywords were needed to make queries as efficient as possible. This paper studied only publicly existing dictionaries / libraries, and did not examine non-publicly accessible concept libraries which have a restricted accessibility through the network of the hosting organizations / institutes.

Conclusion

The seven libraries identified have been developed independently and appear to replicate in different ways similar concepts. Collaboration between similar libraries would greatly facilitate the use of these libraries for others. The process of building code lists takes time and effort. Access to existing code lists increases consistency and accuracy of definitions across studies. Concept library developers should collaborate with each other to raise awareness of their existence and of their various functions, which could increase users' contributions to those libraries and promote their wide-ranging adoption.

Acknowledgments

Kuwait Cultural Office in London, HDRUK and the National Centre for Population Health and Wellbeing supported this research.

Conflicts of interest

The authors declare they have no conflicts of interest.

Ethics statement

Ethical approval to conduct the research was approved was provided by the Research Ethics Sub-Committee, Swansea University.

References

1. Paraskevas Vezyridis ST. Open access research evolution of primary care databases in UK: a scientometric analysis of research output. BMJ Open [Internet]. 2016; Available from: https://doi.org/10.1136/bmjopen-2016-012785
2. SAIL databank. Increase innovation with data linkage. Retrieved January 28, 2020, from https://saildatabank.com
3. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, et al. Defining disease phenotypes using national linked electronic health records: A Case study of atrial fibrillation. 2014; 9. https://doi.org/10.1371%2Fjournal.pone.0110900
4. Lu M, Rupp LB, Trudeau S, Gordon SC. Validity of an automated algorithm using diagnosis and procedure codes to identify decompensated cirrhosis using electronic health records. 2017; 369–76. https://doi.org/10.2147/CLEP.S136134
5. Makarov A, Kontopantelis E, Sperrin M, Stocks SJ, Williams R, Rodgers S, et al. Primary care medication safety surveillance with integrated primary and secondary care electronic health records: A cross-sectional study. Drug Saf. 2015; 38(7):671–82.
6. Nicholson A, Tate AR, Koeling R CJ. What does validation of cases in electronic record databases mean? The potential contribution of free text. Wiley Online Libr. 2011 ;(Ci):321–4. https//doi.org/10.1002/pds.2086
7. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. J Am Med Informatics Assoc. 2013; 20(E1):147–54. https://doi.org/10.1136%2Famiajnl-2012-000896
8. Pendergrass SA, Crawford DC. Using electronic health records to generate phenotypes for research. Curr Protoc Hum Genet. 2019; 100(1):1–20. https://doi.org/10.1002%2Fcphg.80
9. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. 2014; 9(6):6–11. https://dx.doi.org/10.1371%2Fjournal.pone.0099825

10. Bhattarai N, Charlton J, Rudi sill C, Gulliford MC. Coding, recording and incidence of different forms of coronary heart disease in primary care. PLoS One. 2012; 7(1). https://doi.org/10.1371%2Fjournal.pone.0029776

11. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PLoS One. 2009; 4(9): e7168. https://dx.doi.org/10.1371%2Fjournal.pone.0007168

12. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: A case study of atrial fibrillation. PLoS One. 2014; 9(11). https://dx.doi.org/10.1371%2Fjournal.pone.0110900

13. Ostapyk T. Manitoba Centre for Health Policy Data Repository. In: Michalos AC (eds) Encyclopedia of quality of life and well-being research. Springer, Dordrecht; 2014. https://doi.org/10.1007/978-94-007-0753-5_3483

14. Arksey H, O'Malley L. Scoping studies: Towards a methodological framework. Int J Soc Res Methodol Theory Pract. 2005; 8(1):19–32. https://doi.org/10.1080/1364557032000119616

15. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's database of genotypes and phenotypes: DbGaP. Nucleic Acids Res. 2014; 42(D1):975–9. https://doi.org/10.1093%2Fnar%2Fgkt1211

16. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012; 41(6):1625–38. https://doi.org/10.1093%2Fije%2Fdys188

17. RD-Connect. Bioinformatic tools. Retrieved from: https://rd-connect.eu/what-we-do/bioinformatic-tools/

18. Kirby JC, Speltz P, Rasmussen L V., Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Informatics Assoc. 2016; 23(6):1046–52. https://doi.org/10.1093%2Fjamia%2Focv202

19. University of Manchester Institute of Population Health. Clinicalcodes.org (n.d. para. 4). Clinicalcodes.org. Retrieved January 28, 2020, from https://clinicalcodes.rss.mhs.man.ac.uk/

20. University of Manitoba. (n.d. para. 1, 2). Concept dictionary and glossary for population based research. Retrieved January 29, 2020, from http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/concept_dictionary.html

21. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019; 26(12):1545–59. https://doi.org/10.1093%2Fjamia%2Focz105

22. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019; 35(22):4851–3. https://doi.org/10.1093%2Fbioinformatics%2Fbtz469

23. Information Engineering Branch, National Centre for Biotechnology Information, USA. (n.d. para.1). dbGaP Overview. Retrieved January 30, 2020, from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html

24. RD-Connect. (n.d. para 1). RD-Connect. Retrieved from https://platform.rd-connect.eu/

25. Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014; 29(SUPPL. 3):780–7. https://doi.org/10.1007%2Fs11606-014-2908-8

26. Smith, M1, Turner, K1, Bond, R1, Kawakami, T1, and Roos L. The Concept Dictionary and Glossary at MCHP: Tools and techniques to support a population research data repository. 2019; 0(December):1–4. https://doi.org/10.23889/ijpds.v4i1.1124

27. Shah A. CALIBERcodelists user guide. 2014; 1–23. Available from: https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/CALIBERcodelists/inst/doc/userguide.pdf?root=caliberanalysis

28. Cardiovascular Epidemiology Unit, University of Cambridge. (n.d para 3.). PhenoScanner V2, Licence. Retrieved from http://www.phenoscanner.medschl.cam.ac.uk/about/

29. University of Manitoba. (n.d. para. 1). Concept Dictionary and Glossary for Population Based Research. Retrieved January 31, 2020, from http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/concept_dictionary.html

30. RD-Connect. (n.d. para 1, 4). Genome-Phenome Analysis Platform. Retrieved January 31, 2020, from https://rd-connect.eu/what-we-do/omics/gpap/

31. Wong KM, Langlais K, Tobias GS, et al. The dbGaP data browser: A new tool for browsing dbGaP controlled-access genomic data. Nucleic Acids Res. 2017;45(D1):D819-D826. https://dx.doi.org/10.1093/nar/gkw1139

Abbreviations

CALIBER	Clinical disease research using Linked Bespoke studies and Electronic health Records
CPRD	Clinical Practice Research Datalink
EHR	Electronic health records
GPAP	Genome-Phenome Analysis Platform
HES	Hospital Episode Statistics
ICD	International Classification of Disease
MCHP	Manitoba Centre for Health Policy
NLP	Natural Language Processing
PEDW	Patient Episode Database for Wales
PPV	Positive Predictive Value
PheKB	Phenotype knowledgebase
SAIL	Secured Anonymised Information Linkage
SMR	Scottish Morbidity Records
THIN	The Health Improvement Network
UCL	University College London
dbGaP	Genotypes and Phenotypes Database

4 CHAPTER 4: CONCEPT LIBRARIES FOR AUTOMATIC ELECTRONIC HEALTH RECORD BASED PHENOTYPING: A REVIEW

4.1 INTRODUCTION

Electronic health records (EHRs) have been adopted across the UK. For example, in terms of primary care in the UK there are the following four databases: 1) CPRD (Clinical Practice Research Data Link) [8] 2) The Health Improvement Network (THIN) [9] 3) QResearch [10], 4) SAIL (Secured Anonymized Information Linkage) in Wales [11]. In addition, secondary care data such as the hospital admission system (HES-England, PEDW-Wales, SMR-Scotland), are linked to primary care records [6] [7] [13] [14]. Such linked information creates the opportunity to undertake research into the causes and outcomes and pathway of disease. However, using linked routine data requires some specialist skills, for example, using the data requires: 1) identifying conditions of interest from diagnosis, treatments, and procedures, and 2) creating phenotype algorithms (such as diagnosis of rheumatoid arthritis and medication for rheumatoid arthritis) and developing specific inclusion and exclusion criteria [87].

The construction of phenotype algorithms enables repeatable research and ensures that different researchers are using the same standards to identify patients [89]. However, the process of constructing phenotype algorithms is very time consuming and resource costly [142], and so reusing previously created phenotype algorithms to conduct repeatable research becomes a compelling solution. However, it is not common for researchers to share their clinical code lists in their published studies [3]. Therefore, it is difficult to make comparisons between studies as different studies often have different definitions of the same condition [14].

Although clinical code lists were published along with some EHRs based studies, researchers often find it difficult to extract the relevant parts from lists for other research studies. Consequently, it is difficult to evaluate the transparency of EHRs based research [143]. Even though researchers request better transparency in publishing clinical code lists [91] [92], currently journals and funding parties do not make it mandatory to publish code lists [143].

To address these challenges and ensure scientific transparency, data linkage centres have developed concept libraries for disease phenotyping, working as platforms to enable storing, managing, and sharing of phenotypes (Diagnoses, Symptoms, Medications and Procedures) by

In 2012, the government of the United Kingdom established four new research centres for linked electronic health records in London, Manchester, Dundee, and Swansea. Because concept libraries use the linked electronic data from these centres, I limited my search to 2010 or after. This restriction may make it difficult to identify concept libraries created before 2010 in other countries.

multiple researchers. For example, ClinicalCodes.org and CALIBER in the UK, and The Concept Dictionary and Glossary in Canada [96] [143] [144]. In the literature, concept libraries for disease phenotyping have different names and various definitions. We aim to review the literature of existing concept libraries to examine how they are used, identify the current gaps and future development. By evaluating the existing concept libraries and scoping what is missing in the current environment, this study could facilitate the development and improvement of concept libraries.

4.2 METHODS

The five-stage framework of Arksey and O'Malley was used for the literature search [145]. This approach included defining the research questions, identifying relevant studies through literature review, selecting qualified studies, charting and collecting data, and summarising and reporting the findings.

4.2.1 Defining the research questions

This stage involved identifying the research questions, which provided the roadmap for subsequent stages. The questions to be addressed were:

- What concept libraries already exist?
- What are their features? Are there similarities or differences among them?

4.2.2 Identification of relevant studies

This stage involved identifying the relevant studies and developing a decision plan for where to search, which terms to use, which sources are to be searched, time span, and language. Searching was limited to peer reviewed manuscripts which were written in the English language and were published from 2010 to 2019. In 2012, the government of the United Kingdom established four new research centres for linked electronic health records in London, Manchester, Dundee, and Swansea. Because concept libraries use the linked electronic data

from these centres, I limited my search to 2010 or after. This restriction may make it difficult to identify concept libraries created before 2010 in other countries. Five databases were searched including Medline, CINAHL, LISTA, Google Scholar, and Web of Science using the following sets of key words:

1. "electronic health record*" or "electronic medical record*" or "computerized health record*" or "computerized medical record*" or EHR or EMR
2. portal* or platform* or repositor* or library* or dictionary*
3. phenotyp* or e-phenotyp* or phenomic* OR "clinical code list*" or "clinical code*" or "clinical concept*" OR "clinical code set*" or "clinical value set*"
4. The sets of key words have been altered to be used in Google Scholar as recommended by this database as follows: ("electronic health record*" or "electronic medical record*" or EHR or EMR) AND (phenotyp*) AND (portal* or platform* or repository* or library* or dictionary*)

4.2.3 Selecting of eligible studies

I reviewed all the abstracts of the identified manuscripts (N=239) based on their relevance to the research questions. Those with relevant abstracts were taken forward to full assessment (n=50). Out of the fifty fully assessed manuscripts, only seven were selected as they matched the planned inclusion and exclusion criteria. The inclusion criteria for the selection process were to include manuscripts about public concept libraries for electronic linked health data-based phenotyping, and their different definitions, types, and functions, such as allowing users to share, reuse, and verify research methods (e.g., code lists, algorithms, and metadata). Manuscripts related to electronic health record phenotyping authoring tools are excluded. Figure 4.1 depicts more information about the selection process of the related studies, and Table 4.1 presents an overview of the seven concept libraries including their definitions and purposes, electronic data sources, coding systems, and examples of phenotype definitions in the seven public concept libraries.

Table 4.1: An overview of the seven concept libraries

Concept Libraries	Definitions/Purposes	Developers/Leaders	References of the Manuscripts/URL Access of the Concept Libraries	Electronic data sources/Coding systems	Examples of phenotypes
1. General libraries					
ClinicalCodes.org	An online repository that contains a set of published studies. For each study a code list or a group of code lists has been uploaded on the ClinicalCodes.org site. Code lists are publicly accessible to improve validity and reproducibility of electronic medical record studies.	The University of Manchester. Institute of Population Health, UK	Spring ate DA, Ketopantoic E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. 2014; 9(6):6–11. https://clinicalcodes.rss.mhs.man.ac.uk/	Primary and secondary care using Read, OXMIS, SNOMED, CPRD, product/medical code, BNF code, ICD-9, ICD-10	Research article: Are symptoms of insomnia in primary care associated with subsequent onset of dementia? A matched retrospective case-control study, Link to the shared phenotypic descriptions at: https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/article/78/
CALIBER research platform	An open online repository of phenotyping algorithms that contains all definitions of research variables using CALIBER data sources in order to encourage research and promote transparency.	Led from the University College London (UCL) Institute of Health Informatics, UK	Dewaxes S, Gonzalez-Inquired A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J	Primary care, hospital records, social deprivation information, cause-specific mortality data. Using Read codes (a subset of SNOMED-CT), ICD-9, ICD-10, OPCS-4 (analogous to Current	Abdominal Hernia: "At the specified date, a patient is defined as having had Abdominal Hernia If they meet the criteria for any of the following on or before the specified date. The earliest date on which the individual meets any of the

	Am Med Inform Assoc. 2019; 26(12):1545–59. https://www.caliberresearch.org/portal/phenotypes	Procedural Terminology terms) and Gemscript.	following criteria on or before the specified date is defined as the first event date: Primary care 1. Abdominal Hernia diagnosis or history of diagnosis or procedure during a consultation OR Secondary care 1. ALL diagnoses of Abdominal Hernia or history of diagnosis during a hospitalization OR Secondary care (OPCS4) 1. ALL procedures for Abdominal Hernia during a hospitalization" Link to the shared phenotypic descriptions at: https://www.caliberresearch.org/portal/phenotypes/chronological-map

The MCHP Concept Dictionary and Glossary	The Concept Dictionary includes comprehensive operational definitions and programming code for measurements used in MCHP research including a description of the problem(s) involved, methods used, and programming tips/ cautions, and the Glossary records terms that are widely used in research based on population. The Concept Dictionary was developed to help researchers use reliable, validated algorithms to perform methodologically comprehensive research.	The Manitoba Centre for Health Policy (MCHP), Canada	Ostapyk T. Manitoba Centre for Health Policy Data Repository. In: Michalos AC (eds) Encyclopaedia of quality of life and well-being research. Springer, Dordrecht; 2014. http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/concept_dictionary.html	- The MCHP databases: Health, Education, Social, Justice, Registries, Support Files. - Operational definitions and SAS program code for variables or measures developed from administrative data. - The International Classification of Disease (ICD) diagnoses or ICD / CCI (Canadian Classification of Health Interventions) procedure / intervention	Manitoba Asthma Algorithms The following is an example of asthma algorithm developed by a research project. “Raymond et al. (2011) use a broader scope in their definition for asthma, defining it as one physician claim OR one hospital claim with a corresponding diagnosis of: ICD-9-CM: 464, 466, 490, 491, 493 or ICD-10-CA: J04, J05, J20, J21, J40, J41, J42, J45, J441, J448 OR one prescription for an asthma medication in a three-year period. “ Link to the shared phenotypic descriptions at: http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1305#a_references
Phenotype knowledgebase (PheKB)	An online environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms. The	Led by Vanderbilt University, (the eMERGE Network Coordinating Centre), USA	Kirby JC, Speltz P, Rasmussen L V., Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalogue and workflow	Clinical and genomic data from electronic health records.	Urinary Incontinence The cohort is defined with the following criteria:

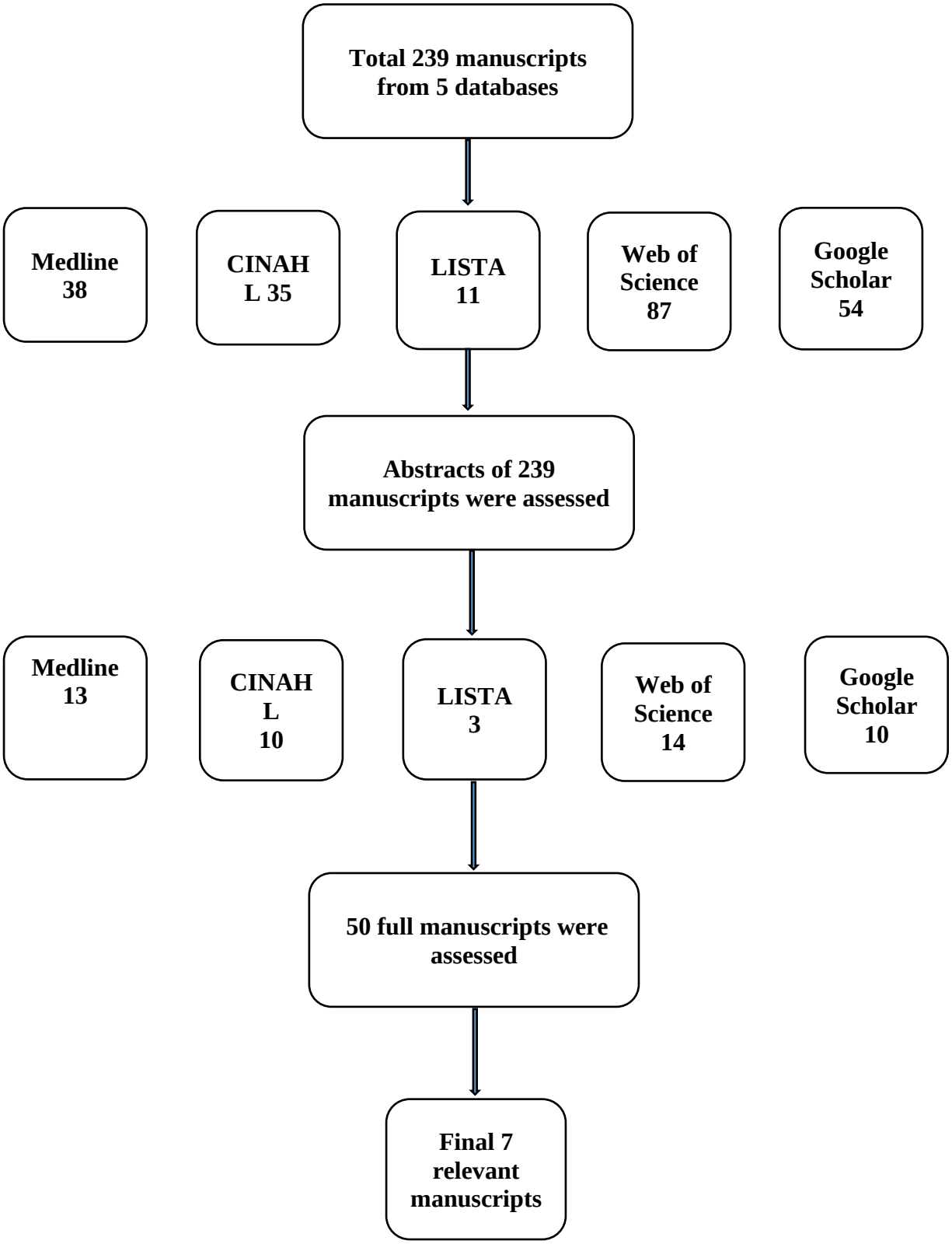
	PheKB was designed to facilitate the transportability of algorithms into various research applications across different organizations, health care systems, and repositories of clinical data.	for creating electronic phenotype algorithms for transportability. J Am Med Informatics Assoc. 2016; 23(6):1046–52. https://phekb.org/	HCPT Codes, ICD 10 Codes, ICD 9 Codes, Laboratories, Medications, Natural Language Processing	a. EHR of all male patients of 35 years of age or more, AND b. For which there is an ICD-9-CM / ICD-10-CM diagnosis of prostate cancer, AND c. For which there are at least two encounters before first treatment, AND d. For which there is at least one clinical note before first treatment, AND e. For which there is either prostatectomy surgery or radiation procedure performed as identified by CPT codes.	Link to the shared

					phenotypic descriptions at: https://phekb.org/phenotype/1404
2. Specialized libraries					
Genome-Phenome Analysis Platform (GPAP)	An online data platform, where data from sequencing experiments contributed by collaborating research projects is processed using a standard pipeline and made accessible to registered users for online analysis through a user-friendly interface.	It was developed by RD-Connect and Led by Aix-Marseille University Medical School (AMU), France	Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-Connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014; 29(SUPPL. 3):780–7. https://dx.doi.org/10.1007%2Fs11606-014-2908-8	Genomic and clinical data from RD-Connect's partners rare disease-based research projects. The PhenoTips database stores phenotypic profiles for individual cases coded by human phenotype ontology (HPO). A directory of biobanks and patient registries and a bio sample catalogue.	Case 1: description RD-Connect identifier: Case1C Gender: Male, Age: 5 years, Referral: Congenital myasthenic syndrome, Onset: Congenital, Global pace of progression: Progressive (slow), Main clinical features: Neonatal hypotonia, Distal arthrogryposis, Inability to walk, Recurrent lower respiratory tract infections. Link to the shared phenotypic descriptions at: https://playground.rd-connect.eu/
The PhenoScanner V2	A database that contains publicly existing results of large-scale genomic association studies. It was	The Cardiovascular Epidemiology Unit,	Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an	137 genotype–phenotype association datasets, including results for anthropometric traits,	Trait: Crohn's disease "A gastrointestinal disorder characterized by

	developed to facilitate the cross-referencing of genetic variants with a wide variety of phenotypes for better comprehension of biology and pathways of disease.	University of Cambridge, UK	expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019; 35(22):4851–3. http://www.phenoscanner.medschl.cam.ac.uk/	blood pressure, lipids, cardiometabolic diseases, renal function measures, glycemic traits, inflammatory diseases, psychiatric diseases and smoking phenotypes. It also includes the NHGRI-EBI GWAS catalogue, and dbGaP catalogues of associations.	chronic inflammation involving all layers of the intestinal wall, noncaseating granulomas affecting the intestinal wall and regional lymph nodes, and transmural fibrosis. Crohn disease most commonly involves the terminal ileum; the colon is the second most common site of involvement. A chronic transmural inflammation that may involve any part of the DIGESTIVE TRACT from MOUTH to ANUS, mostly found in the ILEUM, the CECUM, and the COLON. In Crohn disease, the inflammation, extending through the intestinal wall from the MUCOSA to the serosa, is characteristically asymmetric and segmental. Epithelioid GRANULOMAS may be seen in some patients." Link to the shared phenotypic descriptions at: https://www.ebi.ac.uk/gwas/efotraits/EFO_0000384

Genotypes and Phenotypes Database (dbGap)	A National Institute of Health-sponsored repository tasked with archiving, curating and distributing information provided by studies examining genotype and phenotype interactions. It was developed with standardized identifiers that allows published studies to address or cite the primary data in a clear and uniform way.	National Centre for Biotechnology Information, USA	Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's database of genotypes and phenotypes: DbGaP. Nucleic Acids Res. 2014; 42(D1):975–9. https://www.ncbi.nlm.nih.gov/gap/	Genetic and phenotypic databases sponsored by NIH and other agencies around the world including: Genotype, phenotype, exposure, expression array, epigenomic and pedigree data from genome-wide association studies (GWAS), sequencing studies and other large-scale genomic studies.	"Autism_Genome_Project _Subject_Phenotypes: The subject phenotype table includes data collected on sociodemography (n=2 variables; sex and European ancestry) and psychological and psychiatric observations (n=8 variables; spectrum and strict definition of autism, whether the subject is non-verbal and/or verbal, has low or high IQ, and the age of their first word and phrase). This table now also includes the stage of the study in which the individual was present and the whether individual is a member of a multiplex or simplex family." Link to the shared phenotypic descriptions at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/variable.cgi?study_id=phs000267.v5.p2&phv=161303&phd=3659&pha=3690&pht=2305&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1

Figure 4.1 The selection process of the related studies



4.2.4 Extraction, charting, and synthesis of data

Data was extracted from the seven related manuscripts using a data-charting form. A narrative review method was used to extract data about the investigated seven public concept libraries for electronic linked health data-based phenotyping including their names, types, and characteristics such as enabling users to share, validate, and reuse of research methods such as algorithms.

4.2.5 Collecting, summarising and reporting the findings

A thematic construction was used to provide an overview of the breadth of the literature, and then a thematic analysis was used to generate the results. The different types and the characteristics of the seven public concept libraries were summarised. The types of electronic data sources used in each library (e.g., primary or secondary care or genetic data) and the used coding system (e.g., Read, OXMIS, ICD-9, and ICD-10) were all reported.

4.3 RESULT

4.3.1 The identified public concept libraries from the literature

There were seven public concept libraries from the literature developed by different countries including UK, USA, and Canada. These libraries were the ClinicalCodes.org [143], the Genotypes and Phenotypes Database (dbGaP) [146], Phenotype knowledgebase (PheKB) [147], the Manitoba Centre for Health Policy (MCHP) Concept Dictionary and Glossary [148], Clinical Disease Research using Linked Bespoke Studies and Electronic Health Records (CALIBER) [19], the PhenoScanner V2 [149], and The Genome-Phenome Analysis Platform (GPAP) [150]. Four of the libraries were general libraries and held concepts and phenotypes on multiple specialities ranging from specific conditions (such as codes to identify lupus) to general demographic concepts (such as smoking status). Three of the libraries were specialised libraries and only give concepts on certain defined areas such as rare diseases. However, in common across all the libraries was that they allowed users to share by uploading their own concepts, to examine validity of concepts by showing where they were published, and all had the aim of facilitating reuse of research methods such as clinical code lists or metadata.

There were some clear differences between the general libraries such as archiving data from different electronic sources (e. g. primary care, secondary, social deprivation information, cause-specific mortality data, health, education, justice, and registries); using various types of coding systems (e.g. SNOMED, BNF, READ, ICD9 /10, and Canadian Classification of Health Interventions (CCI) [143] [147] [148] [19]; having different policies that govern accessing the underlying data sources (e.g. a researcher has to complete the Data Access Process (DAP) of the Manitoba Centre for Health Policy (MCHP) to access the data and conduct research by using the Manitoba Population Research Data Repository) [148]; and allowing different searching queries such as simple or more advanced searches (e.g. CALIBERcodelists package enables users to search for code lists by synonym or code stub, and combine search terms using Boolean operators) [151].

However, they share some clear similarities such as having similar purposes (e.g. helping researchers to perform comprehensive research, promoting transparency in sharing research methods, and improving reproducibility of studies; enabling users to upload their own code lists and other important documents (e.g. users of PheKB can upload related documents and their phenotypes along with multidimensional metadata labels, documents including detailed descriptions of the computable algorithms such as types of used data, logic of execution,

definitions of data, and flow charts) [147]; and allowing users to use/download the publicly accessible code lists (e.g. users of ClinicalCodes.org can download a file containing all codes associated with a study as csv files) [143].

There were some differences between the specialized libraries such as using data that were generated by various electronic databases (e.g. the PhenoTips database, database of biobank and patient registries, the NHGRI-EBI GWAS catalogue, and genetic and phenotypic databases sponsored by NIH and other agencies around the world) [146] [149] [152]; allowing diverse searching strategies (e.g. registered users of GPAP may select one or more individuals such as trios or other family relationships to explore and then filter and refine the outcomes by inheritance mode, population frequencies, tools for in silico pathogenicity prediction, gene lists and Linvar, HPO and OMIM codes) [153]; and enabling different searching queries such as simple or complex searches (e.g. all publicly released dbGaP studies can be queried by users. Queries can be very simple, just a keyword of interest ('cancer') or complex, making use of search fields and Boolean operators ('cholesterol[variable] AND phs000001') [146].

Conversely, there were some similarities between the specialised libraries such as enabling users to share by uploading their own concepts in the libraries to analyse the validity of concepts by showing where they were published (e.g. the GPAP enables clinicians and researchers who upload patient datasets to analyse their own data [153], and NIH-funded researchers can share their produced data, in the dB Gap) [146]; all aimed to encourage the reuse of research methods such as lists of clinical code or metadata (e.g. registered users of GPAP are allowed to access and search data sets provided by other researchers on similar patients [153], and users of the PhenoScanner V2 can use the archived findings from large-scale genetic association studies which are publicly accessible) [149]; and allowing access to some datasets through specific established control access (e.g. individual level data is accessible in the dbGap to scientists around the world through controlled application of access) [146]. Information about all the seven concept libraries such as their access URL and references of the seven manuscripts are presented in Table 4.1.

4.3.2 Concept libraries names and definitions:

Each of the investigated concept libraries has a specific name and a unique definition (Table 4.2). For example, CALIBER is defined as "a unique research platform consisting of 'research ready' variables extracted from linked electronic health records (EHRs) from primary care, coded hospital records, social deprivation information and cause-specific

mortality data in England" (https://www.ucl.ac.uk/health-informatics/caliber). Whereas, the Database of Genotypes and Phenotypes (dbGap) is defined as “a National Institutes of Health-sponsored repository charged to archive, curate and distribute information produced by studies investigating the interaction of genotype and phenotype" [146].

4.3.3 Concept libraries types

The explored concept libraries (N = 7) differ in their types. Some are general libraries holding several specialist phenotypes (n = 4) such as ClinicalCodes.org, which holds code lists for all published electronic medical record studies, irrespective of code types (such as Read, ICD9-10, SNOMED) that are collected from multiple databases (such as CPRD, Research, THIN), while others are specialised libraries which only manage certain specificities (n = 3) such as the Genome-Phenome Analysis Platform (GPAP), which holds genomic and clinical data from rare disease based research projects.

4.3.3.1 General concept libraries:

1. **The ClinicalCodes online repository**

The ClinicalCodes repository contains a selection of published studies that have been uploaded to the ClinicalCodes.org site along with a code list or a series of code lists. A code name, coding system (Read, OXMIS, SNOMED, CPRD product / medical code, BNF code, ICD-9, ICD-10), definition and type of entity (diagnostic, drug, examination, clinical sign, administrative, demographic, observational, immunization) are assigned to all individual clinical codes. Metadata and links to studies code lists are accessible as research objects that could be shared in machine-readable form throughout platforms. A research object file of JavaScript Object Notation (JSON) is available for each study that contains metadata (title, author, abstract, reference, link, DOI), commentary on the study level, commentary on the code list level and links to the individual files of the code list. Such object research files are directly accessible when inserting a '/ro' to the URI for a study e.g., (www.clinicalcodes.org/medcodes/article/5/ro) [143]. The developers of the ClinicalCodes repository have created an open-source R package (rClinicalCodes) to automate the downloading and importing lists of clinical code and metadata through the research object file from the repository website: (https://cran.r-project.org/web/packages/rClinicalCodes/index.html. The developers of the ClinicalCodes repository will implement in the future: 1) Searching and downloading of codes by disease

group, keyword and/or code group. 2) Methods for downloading code-lists and article metadata in machine readable form. 3) An API for downloading code-lists programmatically [143].

2. **The Clinical Research using Linked Bespoke Studies and Electronic Health Records (CALIBER)**

CALIBER has developed the CALIBERcodelists package to manage ICD-10, Read and OPCS coding lists to identify medical conditions for research using CALIBER or other UK electronic health record databases. The package is written in R language, but many of the functions are accessible through an interactive menu and do not require any experience with R. The package has many features: 1) provides a standardized approach to identify codes of interest including Read, ICD-10 and OPCS through searching for term text or codes, 2) enables displaying of code lists on a spreadsheet and removing individual terms or modifying their categories, 3) downloads code lists in a variety of formats, and uploads them in default file format, 4) allows comparing of one code list to another, or combine two code lists together, 5) enables converting of code lists across dictionaries using the NHS mapping between the terminologies of Read/OPCS and Read/ICD-10, 6) processes a document which contains code to produce a code list and a descriptive text, and produces a comprehensive HTML document and a standardised format code list [151].

3. **The Manitoba Centre for Health Policy (MCHP) Concept Dictionary and Glossary**

MCHP has built a range of web-based tools that record the historical usage of the repository-saved information such as the MCHP Concept Dictionary and Glossary [148]. Although there are short definitions for widely used terminology in the glossary, the Concept Dictionary includes comprehensive operational definitions and programming code for measurements used in MCHP research. A coherent documentation approach is used to describe the research methodologies. They can be presented either as best practices, or as different versions of historical overviews. Enhancements substitute older versions with a best-practice approach which requires authoritative approval of what is "the best", and the historical overview records all published methods for making options accessible to the user and information about the methodologies used in previous studies [154].

4. **The Phenotype Knowledgebase (PheKB)**

The Phenotype Knowledgebase (PheKB accessible at http://phekb.org) was created within the eMERGE Network as a workflow management system and learning centre to support computable algorithm creation, validation, and sharing. It enables the transportability of algorithms across various research applications, multiple organizations, health care systems, and clinical data repositories through feedback processes and standardised implementation performance measures. PheKB contains built-in tools designed specifically to improve sharing of knowledge across sites, for example, the Data Dictionary / Data Validation Tool and the data management function. The Data Dictionary / Data Validation Tool is a registered user embedded resource that validates definitions of covariate data and related data, promotes data standardization, and early-stage quality assurance to exchange data for study sets efficiently. It is used for identifying errors and warnings in data dictionaries and data files related to a given phenotype through a custom Drupal module. It can show errors and warnings regarding the structure and content of the files as files are uploaded, while the data management function provides tracking tools for users to easily determine what data has been shared, what algorithm it is linked to, and by whom it was shared [147].

4.3.3.2 The specialized concept libraries:

1. **The Genome-phenome analysis platform (GPAP)**

The RD-Connect built an integrated Genome-phenome analysis platform (GPAP), which is a user-friendly tool for diagnosing and discovering genes [153]. It links anonymised omics and clinical data with tools and services to examine these data online. The main portal provides links to the genomics analysis interface and the Phenotypes database that store ontology of phenotypic profiles coded for individual cases by Human Phenotype Ontology (HPO). GPAP also includes a database of biobanks and patient registries, and a catalogue of bio samples that allows information of individual samples housed in participating biobanks to be drilled down [155]. For example, a researcher may select one or more individuals (e.g., trios or other family relationships) to explore and then filter and refine the outcomes by inheritance mode, population frequencies, tools for in silico pathogenicity prediction, gene lists and Linvar, HPO and OMIM codes [152] [153].

2. The PhenoScanner V2

The developers of PhenoScanner V1 have collected more than 5,000 genotype-phenotype association datasets to create version 2 of the catalogue (PhenoScanner V2). PhenoScanner V2 has an API that contains an R package and a Python command line tool associated with it, which enables users to search for PhenoScanner V2 genotype-phenotype associations within R or from a terminal. All results, irrespective of P-value, can be presented when querying genetic variants, allowing the user to find indication against phenotype associations. PhenoScanner V2 has new features to facilitate improved 'phenome scans' including: 1) an expanded database of human genotype-phenotype associations divided into phenotype classes (diseases and traits, gene expression, proteins, metabolites and epigenetics), 2) new search selections such as gene, genomic region and queries based on phenotypes 3) Linkage Disequilibrium (LD) information for the five super-ancestries in 1000 Genomes 4) variant annotation and trait ontology mappings 4) annotation variations and ontology mappings of traits, and 5) a new Platform and API [149].

3. The database of Genotypes and Phenotypes (dbGaP)

The database of Genotypes and Phenotypes (dbGaP) enables licensed users to identify and display different regions of the human genome, such as all the Allele frequencies and subgroups of individual-level genotype as well as sequence data, which are stored in that region in dbGaP, without accessing data sets of interest and performing multiple analyses [156]. The browser uses the standard graphical interface built for data from the 1000 Genomes Project and dbGaP by the National Centre for Biotechnology Information (NCBI), which incorporates sequence viewer track views with genotype tables and a novel sample / subject data selector showing core sample phenotype data [146]. The webpage of the browser includes a selection of 'widgets' pages showing data from the dbGaP view-only data project, which is data from the collection of dbGaP General Research Usage (GRU). The widgets work in such a way that one widget operation causes updating of other widgets on the page. See online browser documentation (https:/www.ncbi.nlm.nih.gov/gap/ddb/help/) and The NCBI YouTube channel (https: /www.youtube.com/user/ NCBINLM) for additional widget information.

4.3.4 Concept libraries characteristics

Some of the characteristics of the seven concept libraries are presented in Table 4.2.

4.3.4.1 Sharing of Concepts

All of the identified concept libraries allow researchers to share some or all of their research methods such as clinical codes list, metadata, and algorithms. For example, users of the ClinicalCodes.Org are able to upload the code list and metadata for specific codes, and add comments at the code list, or study level. However, an account should be created first [143]. According to the developers of ClinicalCodes.Org, "To date: 93375 clinical codes have been deposited over 521 code lists" [157]. Similarly, Phenotype knowledgebase (PheKB) enables researchers to upload related documents and their phenotypes along with multidimensional metadata labels including the methods used in the phenotype standards such as International Classification of Disease (ICD) codes, medications, and Natural Language Processing (NLP). Researchers also can upload documents that include detailed descriptions of the computable algorithms, such as types of data used, logic of execution, definitions of data, and flow charts [147].

Builders of the Concept Dictionary and Glossary at the Manitoba Centre for Health Policy (MCHP) encourage researchers to share their discoveries, such as creating new concepts or updating existing concepts to grow and improve the value of this publicly accessible resource. Their Concept Dictionary describes more than 300 research concepts developed at MCHP for the analysis of data contained in the data warehouse hosted at MCHP [148]. Also, the developers of the CALIBER research platform promote collaborative research. They compiled more than 90,000 terms from five standardised clinical terminologies to construct 51 validated phenotyping algorithms (35 diseases or syndromes, 10 biomarkers, 6 risk factors for lifestyles) [19]. All data sources are made accessible to researchers and can be accessed in a secure data-safe haven environment located at UCL IHI / Farr Institute, London or could be accessed remotely. Due to the varied clinical backgrounds of the datasets, they offer training on data sources, coding, consistency and management with the CALIBER team [18].

The Phenoscanner V2 database includes more than 5000 genetic association datasets from publicly accessible datasets of complete summary of associations findings collected by the NHGRI-EBI (https:// www.ebi.ac.uk/gwas/downloads/summary-statistics) and NHLBI (https://grasp.nhlbi.nih.gov/FullResults.aspx), and recent literature reviews and GWAS omics datasets [149]. Also, the Genotypes and Phenotypes Database (dbGaP) allows sharing of information obtained from studies

examining genotype and phenotype interactions [146]. These studies include research of genome, medical sequencing, molecular diagnostic assays, and correlation between genotype and non-clinical traits [156]. Similarly, the Genome-Phenome Analysis Platform (GPAP) facilitates data sharing as it now opens for submissions of projects from all users, and not only from RD-Connect partners [150]. One of their main objectives is to help the contributed projects to quickly make their data available to the broader community of rare disease researchers [152].

4.3.4.2 Validation of Concepts

Most of the identified concept libraries from the literature have described their validation methods either in their published studies, or in their websites, or in both of them. For example, in the CALIBER research platform, EHR-derived phenotypes have been extensively validated using six different approaches: cross-EHR source concordance, case note review, consistency of risk factor-disease association from non-EHR studies, consistency with prior prognosis research, consistency of genetic association, and external populations. The builder of the platform acknowledged that the case study would inform which validation(s) are most important. For example, phenotyping algorithms developed for disease epidemiology (e.g., screening or disease surveillance) might be designed for higher sensitivity whereas those used in genetic association studies might be designed to maximize Positive Predictive Value (PPV) [19].

The developers of the PheKB have developed the Data Dictionary / Data Validation Tool, which validates covariate data descriptions and related data and is a tool embedded for registered users. The user uploads to the phenotype-related page, and the tool verifies the data dictionary file for compliance with standards and best practices. A specified set of rules defines differences from the standard or guidelines for best practices [147].

The Concept Dictionary and Glossary was built at MCHP to assist researchers to carry out methodologically comprehensive research using consistent, validated algorithms [148]. According to their builders, concepts are written using original ideas and methods developed for MCHP reports, then reviewed and shaped according to common standards by the repository analyst [154]. Similarly, the data submitted including individual genomic and phenotype data, analytical results, general study information,

are subject to quality checks by the database of Genotypes and Phenotypes (dbGaP) staff before the dbGaP information is released publicly [146].

4.3.4.3 Reusing of Concepts

All of the seven concept libraries allow reusing of stored data, clinical code lists and algorithms, however each concept library has established certain terms and searching features for users. For example, any user can download code lists from the ClinicalCodes.org repository. Once deposited, code lists will be freely available, with no login needed to download the codes. In addition, an open-source R package has been developed to automate the downloading of code lists from the online repository [143]. Also, the CALIBER research platform allows reuse of existing lists of codes by researchers. Users can access phenotyping algorithms defining over 90 diseases and metadata. CALIBER has CALIBERcodelists package [151], which enables users to search for code lists by synonym or code stub, allows users to combine search terms using Boolean operators, and supports regular expressions for more advanced search queries. In addition, it allows downloading of the list of codes and some basic metadata for example, the name and version of the code list as a csv file [19].

Algorithms and multiple implementation results can be publicly viewed in the PheKB website when authors designate it as “final”. By using metadata, users can search an algorithm based on inclusion or exclusion of data elements classes, such as diagnosis, author, or keyword. Currently, there are 414 users of PheKB from 52 different institutions. The median of used algorithms per institution is four. As of March 2020, PheKB include 30 public algorithms with 66 executions and 62 non-final algorithms with 83 executions in different stages of development [147].

PhenoScanner V2 is a searchable library of findings from large-scale genetic association studies which are publicly accessible. The database now includes more than 350 million association results and more than 10 million original results genetic variations [149]. The developers of the PhenoScanner V2 specified the terms of use in their website: 1) users should cite both their papers in any publication or presentation 2) users should cite the original paper where the results were obtained, including the references for the linkage disequilibrium statistics and variant & phenotype mappings where used and 3) users should comply with any other terms relating to the data [158].

The Concept Dictionary and Glossary at MCHP describes more than 300 research concepts developed at MCHP for the analysis of data contained in the data warehouse hosted at MCHP [159]. Over time, traffic on the MCHP website has increased. Their analysis software, Deep Log Analyser, recorded more than two million visits in 2018. In addition, the Charlson Comorbidity Index (CCI) including a glossary of terms and concepts has been ranked as the most widely viewed definition in the MCHP Concept Dictionary for many years. Similarly, the Elixhauser Comorbidity Index concept has regularly appeared among the top five most viewed concepts; whereas measures of comorbidity have often been among the most viewed concepts [154].

The database of Genotypes and Phenotypes (dbGaP) provides free access to publicly available information on completed research and studies-related documents. However, individual level data is open to scientists around the globe via controlled access application. This platform allows researchers and clinicians to quickly interpret and compare DNA sequencing data with clinical knowledge, including those who do not have training in bioinformatics. Information in the dbGaP is organized as a hierarchical structure and includes the accessioned objects, phenotypes (as variables and datasets), various molecular assay data, analyses and documents. The dbGaP enables both simple as well as advance searches [146].

4.3.5 Concept libraries limitations

Some of the developers of the concept libraries mentioned some of their limitations as described below:

The ClinicalCodes repository does not offer methods for downloading code-lists and article metadata in machine readable form according to their developers, and is lacking search features needed to facilitate queries such as searching and downloading of codes by disease group, keyword and or code group, all of which are planned to be added in the future. Also, they stated that it does not have a protocol for enforcing quoting of the downloaded code lists. Therefore, it would be difficult to connect code lists from earlier studies [143] .

The developers of CALIBER mentioned that there are some fairly complete measures in CALIBER's data, for instance, 82.6 percent of people with at least one measurement of BMI using the concepts in the library. But some measurements are less comprehensive, for example, only 44.9 percent have at least one total measure of

cholesterol when using the library concepts. They also mentioned that different records in CALIBER can represent the same event or subsequent events at similar points in time. For example, fatal myocardial infarction can be reported in up to four diverse sources that vary in their specificity in diagnosis and precision in timing [19].

The developers of PheKB stated that some algorithms cannot work at a given site as well as at another, and validation is the only way to distinguish poorly performing algorithms. They also mentioned that PheKB does not have programmatic interfaces with some of the networks needed to enable fast exchange of executable phenotyping algorithms [147].

The developers of dbGaP declared that the National Institutes of Health (NIH) policies, such as restricting the context of available data only to researchers who consented to general research use, leads to limiting the related phenotype data to specific demographic and disease status information, reducing the data download capabilities of the browser, and showing a browser watermark that images must not be recorded or published [146] [160].

Table 4.2 Some of the characteristics of the seven concepts libraries

Concept Libraries	Access to the underlying data sources	Sharing/Uploading of Concepts	Reusing/Downloading of Concepts
1.General libraries			
ClinicalCodes.org	In the top menu tab 'Browse published studies,' the user may choose a published study from the list. It then shows all the code lists associated with that study.	Users must register with ClinicalCodes.org (in the menu bar login/signup) and choose 'upload codes.' First, they need to add some metadata of the published study and then they can upload several codes lists as delimited text files into that study. Metadata and links to studies code lists could be shared in a machine-readable form using the available open-source R package (rClinicalCodes).	Code lists are released on ClinicalCodes.org using a Creative Commons Attribution 3.0 Unported License (CC BY 3.0), and a file containing all codes associated with a study can be downloaded and used freely by any user. Downloading individual code lists is a single-click process that does not involve logging in or supplying user information. Users can choose to explore and download some or all of the code lists as csv files.
CALIBER research platform	Access to CPRD linked data on the CALIBER research	If a project proposal for a researcher has been accepted, a	All definitions of research variables that use CALIBER data

	platform complies with the governance policies for data access by the CPRD. Researchers should first sign agreements with UCL to access CPRD data. Non-UCL partners must apply for CPRD to become a CPRD-approved partner and sign a UCL-approved sub-license agreement.	registration with the UCL Identifiable Data Handling Service (IDHS) will be arranged in order to create a new share for the project on the safe haven. The data facilitator at CALIBER will direct researchers through the entire process.	sources are publicly available and can be accessed in human and machine-readable formats. CALIBERcodelists package enable users to search for code lists by synonym or code stub, combine search terms using Boolean operators, and download the list of codes and some basic metadata as a csv file.
The MCHP Concept Dictionary and Glossary	The Data Access Process (DAP) of the Manitoba Centre for Health Policy (MCHP) are the processes that a researcher has to complete to access the data and conduct research using the Manitoba Population Research Data Repository.	Researchers can share their work, such as creating new concepts or updating existing concepts. The concept development guidelines are defined in the Concept Development Template. **Concept Development Template**.	**Concept Dictionary**: More than 300 research concepts developed at MCHP are publicly accessible. **Glossary**: Terms of documentations widely used in population-based research are freely available. **Browse/Search the Concept Dictionary and Glossary**

Phenotype Knowledgebase (PheKB)	Private Phenotypes with "In Development" status, phases of "Testing," or "Validation" are not publicly accessible, which can only be accessed if the user is logged in and the phenotype was shared with the user via one of the two collaborative groups: Owner Group Phenotypes or View Group Phenotypes.	Researchers can upload: Related documents and their phenotypes along with multidimensional metadata labels. Documents including detail descriptions of the computable algorithms, such as types of used data, logic of execution, definitions of data, and flow charts.	Algorithms and multiple implementation results can be publicly viewed in the PheKB website when author designated it as "final". By using metadata, users can search an algorithm based on inclusion or exclusion of data elements classes, such as diagnosis, author, or keyword.
2. Specialized libraries			
Genome-Phenome Analysis Platform (GPAP)	Only approved users who have completed the registration and verification process can access the data stored on the GPAP. Users must be affiliated with a recognized academic institution as accredited clinicians/researchers and must demonstrate their approval of the RD-Connect Code of	Data sharing is open for project submissions from all users, not only from partners of RD-Connect, but they have to register first in the GPAP website. The GPAP enables clinicians and researchers who upload patient datasets to analyse their own data.	Registered users are allowed to access and search data sets provided by other researchers on similar patients. Registered users can match make, find second families, and find patient populations for validation studies.

	Conduct by signing the Adherence Agreement.		
The PhenoScanner V2	Some of the datasets are available for download including: dbSNP 147 with variant annotation from VEP, Linkage disequilibrium statistics from 1000 Genomes and a subset of the processed GWAS datasets, but users should first contact phenoscanner@gmail.com to get an approval.	Users can input one genetic variant, gene, genomic region or trait in the home page text box (www.phenoscanne.medschl.cam.ac.uk) or upload as a tab-delimited text file up to 100 genetic variants, 10 genes or 10 genomic regions.	Users can use the archived findings from large-scale genetic association studies which are publicly accessible. Information provided by project members are unrestrictedly accessible.
Genotypes and Phenotypes Database (dbGaP)	Free access to information on completed studies are open to the public. Individual level data is accessible to scientists around the world through controlled application of access.	NIH-funded researchers can share their produced data. However, studies that are not sponsored by the NIH, individual NIH Institutes and Centres (IC) make judgments about whether non-NIH sponsored data should be accepted.	Open-access data can be accessed online or downloaded without prior authorization or permission from dbGaP. Individual level data download requests are handled through the dbGaP Authorized Access System (dbGaPAA), a web portal that manages request submissions,

		and enables safe high-speed large data download for authorized users.

4.4 DISCUSSION

4.4.1 Statement of main findings

Globally, the development and use of concept libraries could be useful for reusable health studies. A number of data linkage centres around the world have developed different concept libraries to facilitate repeatable research. This paper examined seven concept libraries, and variations in their definitions, names, types, functions, coding systems, and data access restrictions. One of our findings is that these concept libraries have developed independently and so are duplicating work but in slightly different ways.

For wide use of concept libraries, collaboration across data linkage centres is needed to develop common standards that govern and guide these emerging libraries. For example, they should agree on a relatively standard definition/name for concept libraries to enable users to locate them and then use them easily. In addition, builders of concept libraries should cooperate with each other to increase awareness about their existence and their various functions. What one concept library might do that other do not do (e.g., provide SNOMED or BNF code lists or provide definitions for demographic variables such as smoking, BMI algorithms that other concept libraries do not do). Raising awareness of the features in the different libraries could increase the contributions of users to these libraries and accelerate their wide adoptions.

For a comprehensive adoption of concept libraries, their various functions, such as enabling users to share, validate, and reuse concepts (e.g., code lists), and their search features should be assessed by developers, funders, users, and experts to ensure that they meet the needs of various users including researchers, clinicians and data analysts. Since there are two different types of concept libraries 1) general libraries that hold phenotypes of multiple specialties 2) specialised libraries that manage only certain specificity such as rare diseases, users' preferences for the type of concept library types needs to be evaluated (e.g., through interviews, focus group, and surveys) before developing new concept libraries.

4.4.2 Strengths and limitations

This is the first study, to our knowledge, aimed at identifying existing concept libraries, exploring their various characteristics, and examining the current practices in this evolving field. Finding studies about a concept library for electronic health data phenotypes in the literature was challenging as there are a limited number of related studies. Another challenge was the lack of a standard name or definition that describes this kind of library. Therefore, a

range of keywords were needed to make queries as efficient as possible. This paper studied only publicly existing dictionaries / libraries, and did not examine non-publicly accessible concept libraries which have a restricted accessibility through the network of the hosting organizations / institutes.

4.5 CONCLUSION

The seven libraries identified have been developed independently and appear to replicate in different ways similar concepts. Collaboration between similar libraries would greatly facilitate the use of these libraries for others. The process of building code lists takes time and effort. Access to existing code lists increases consistency and accuracy of definitions across studies. Concept library developers should collaborate with each other to raise awareness of their existence and of their various functions, which could increase users' contributions to those libraries and promote their wide-ranging adoption.

Chapter 5

Review 2:

This chapter describes some of the challenges that researchers may have when attempting to identify and locate conditions and their code lists for research purposes. It involves conducting a review of the literature to identify existing classification systems used for identifying children with chronic conditions in routine data sources.

5 CHAPTER 5: CLASSIFICATION SYSTEMS FOR IDENTIFYING CHILDREN WITH CHRONIC CONDITIONS IN ROUTINE DATA SOURCES: A REVIEW

5.1 INTRODUCTION

The prevalence of chronic conditions in children has increased over recent decades [161], with estimates ranging from 10% to 30% [162] [163] [164]. Most of the increases are due to the incidence of asthma, obesity, mental health conditions, and neurodevelopmental disorders [165]. The causes of the increase include social changes, prenatal influences, nutrition and physical activity, environmental exposures [161].

This rise in prevalence has significant financial and organisational implications for health-care planning [166]. Chronic conditions cause severe stress in millions of children and adolescents, increasing their risk of emotional and behavioural issues [164]. These chronic conditions also have physical and economic impacts on children and their families, resulting in vulnerability and a lower quality of life and family stability [167]. The rapid expansion of chronic conditions in children will result in a large number of young adults who have chronic conditions and disabilities and participate less in the community. Policymakers and politicians need valid prevalence data to improve these children's societal participation as they approach adulthood and to prepare for adequate and appropriate services [166]. Valid prevalence data could also be used as an outcome measure in comparing indicators of youth health among countries, and for monitoring and evaluating the effect of services and interventions [168] [169]. The ability to support children with chronic conditions in such a way that they have positive health and educational results should be attainable; however, accurate prevalence data are required to make that possible.

A variety of approaches have been used to assess the prevalence and implications of chronic conditions and health disorders in children, resulting in a wide range of prevalence estimates that are difficult to compare [170]. Van Der Lee et al. conducted a systematic review of all definitions and operationalizations used to estimate the prevalence of chronic health conditions in children, and they found variations in definitions (e.g., whether noncategorical definitions or diagnosis lists are used) and variations in operationalizations of similar definitions, including the used source of information (e.g., parents or adolescents, interviews or medical records) [166].The main causes of diversity are the lack availability of standardised criteria for defining chronic conditions in children. For example, Mokkink et al. defined chronic conditions

in children as “a health care problem lasting at least 3 months that involves frequent hospital admissions, at-home medical care and/or other forms of health care” [171] Van Cleave et al., on the other hand, included the social and demographic components of diseases to Mokkink’s definition, and specified a disease duration of at least 12 months [162]. A clear definition is required in order to get valid and reliable estimates of the prevalence of chronic conditions in children [166].

Interest among clinicians and researchers in multimorbidity (co-occurring physical and mental illnesses) is also increasing. Romano et al. found that the current physical-mental multimorbidity literature was primarily focused on Attention Deficit Hyperactivity Disorder (ADHD), anxiety, and mood disorders in children with epilepsy, asthma, and allergy [172]. Any chronic physical illness in children and adolescents significantly increases the probability of developing mental illness [173]. Researchers have estimated that nearly half of all children and adolescents with a diagnosed physical illness also experience mental health problems [174]. Physical and mental health conditions have a considerable impact on severity and functional impairment in children and adolescents. The specific linkages between mood disorders and inflammatory/immunologic conditions, and also the wider relationships between neurologic disorders that affect and behavioural disorders, have significant implications for future aetiology and therapeutic research [175].

The increasing use of routine healthcare and other administrative datasets provides an ideal opportunity to study the prevalence of chronic conditions and multimorbidity in children on a population level, without the need to conduct large-scale long-term cohort studies. However, in order to do so, researchers need to generate code lists from these datasets in order to identify children with chronic conditions. Building code lists takes a lot of effort, and it often necessitates knowledge of complicated computer languages such as SQL [144]. This means that routine healthcare and other administrative datasets remain inaccessible to many researchers since their usage necessitates specialised programming expertise [176].

The availability of clinical codes in routine healthcare and other administrative dataset-based research is one of the most important objects for reproducible research because researchers, clinicians, and health informatics professionals frequently use them to identify the target population and their specific conditions, a process known as phenotyping or phenotyping algorithms [3] [14].To facilitate reproducible research and assure consistency across studies, many data linkage centres have built concept libraries, which work as platforms for multiple

researchers to store, manage, and share phenotypes (Diagnoses, Symptoms, Medications, and Procedures) along with the phenotyping algorithms, and the code lists such as ClinicalCodes.org [90], CALIBER research platform [144], and the Concept Dictionary at the Manitoba Centre for Health Policy [154].

Clinical classification systems are designed to categorise clinical conditions and procedures in order to enable statistical data analysis across the healthcare system and to give criteria for comparing national and worldwide health statistics [33].To fully utilise administrative data systems, clinical classification systems that can be adapted for use across multiple datasets are required to ensure that all conditions are included. This is especially significant if researchers want to know the burden of all chronic childhood conditions rather than just those that result in hospitalizations or deaths. There are a few well-known coding systems, which are mostly used in the USA studies where analysis of routine data sources, particularly health insurance, is common [40].

Two studies have reviewed different classification systems, one comparing four different systems used in the United States and the other comparing an American system with a British system [177] [178]. Berry et al. assessed existing approaches for identifying children with medical complexity by examining a variety of health data sources, including administrative billing data and self-reported surveys by parents or providers. They analysed and contrasted four examples of diagnosis classification systems that have been used to detect health problems experienced by children with medical complexity, including the complex chronic condition, clinical risk group, chronic condition indicator, and patient medical complexity algorithm, using International Classification of Diseases (ICD) diagnosis and procedure codes [177]. While, Hardelid et al. developed a broader definition of chronic conditions in children than previously published definitions in order to characterise children who died from chronic disorders. They constructed a code list (based on the International Classification of Diseases, 10th Revision (ICD10)) to identify chronic condition indices reported in death certificates and hospital administrative records, and they classified the final list of ICD10 codes for chronic diseases into eight categories [178]. Both had a relatively narrow focus, with Berry et al. looking just at hospital data and Hardelid et al. using a classification system for death certificate data and hospital administrative records in the UK.

The aim of this study was to conduct a review of the literature using systematic methods in order to summarise the clinical classification systems used to identify chronic diseases in

children in routine data sources and other administrative dataset, including information on the coding systems upon which they are based (for example, ICD-10), the groupings used (for example, chapter headings within ICD-10), and, if possible, the specific codes used to identify a specific condition.

5.2 METHODS

Eligibility criteria for the review were determined to identify papers that had aimed to measure chronic diseases in childhood using routine sources of data. These sources could include data on hospital admissions, records of visits to outpatients or emergency departments, general practitioner records, insurance system records, or death certificates.

5.2.1 Eligibility criteria

Specific inclusion criteria were that studies should have:

1. Included data only on children or, if the whole population was included, data must have been presented separately for individuals <25 years old (with an older upper limit included because we were aware that some studies of chronic conditions had defined a "children and young people" group in this way).
2. Been quantitative in nature. That is, they used a routine or administrative data source to identify children with chronic conditions in a population. All quantitative study designs were eligible for inclusion.
3. Been written in English.
4. Specified study characteristics (e.g., PICOS, length of follow-up), reported characteristics (e.g., years considered, language, publication status) used as criteria for eligibility (giving rationale), and given a clear description of the data source.
5. Examined all "chronic illness" in childhood, or multiple specific conditions. For our purposes, they must have looked at more than one condition within the study because studies that just look at one condition may have extensive coding systems for that one condition that would be difficult to replicate in population-based studies of all conditions. It was not feasible for us to do a search that would capture all conditions. As a result, we limited it to studies that examined at least two conditions within the same study, as long as the studies employed some type of routine data to make it useful.
6. Provided details of how they defined the diseases AND/OR a description of how cases were identified.

7. Included a population-based sample. Studies based on clinical samples (in which the condition has been identified based on the fact that they receive care from a specific physician, for example) were excluded. Studies based on self or parental report of chronic conditions were excluded. Studies including a specific group of children (such as children in the welfare system, or criminal justice system) could be included, as long as the data collection on them was population-based (for example, using an administrative dataset to identify all children in this group in a particular area). Studies that screened children in the community and then conducted a study of those identified as having a specific problem and studies that measured the prevalence of symptoms associated with a chronic disease are also excluded.

5.2.2 Search

The Medline database was searched on January 30th 2020. The search strategy included a set of Medical Subject Headings (MeSH) and free text terms to identify studies in children, studies on chronic diseases, and studies including terms for measurement, identification or classification (See table 5.1). Within each category, the terms were combined using OR, and then the three sets of terms were combined using AND. No limits were placed on the search results. Reference lists of included papers and key background references were searched for additional studies.

5.2.3 Study selection

The results of the Medline search were entered into an Endnote library where titles and abstracts were first screened, followed by the creation of a list of studies for full text screening. 10% of the titles and abstracts and full texts were screened independently by two authors. Disagreements were resolved by discussion.

5.2.4 Data collection process

The data extraction sheet was created for this review, which is an excel document with three different sheets: the first sheet was designed to collect the basic characteristics of the study; the second to summarise the main findings of each study; and the third to extract the actual information on the codes (See table 5.5).

Table 5.1: Medline search strategy

	Multimorbidity Medline search	
Searches	**Results**	**Type**
1	child*.mp. or Child/ or Child Health.mp.	2388861
2	paediatric*.mp.	64904
3	pediatric*.mp.	338067
4	Pediatrics/	52630
5	1 or 2 or 3 or 4	2476109
6	Chronic Disease/ or chronic condition.mp.	263147
7	chronic conditions.mp.	13899
8	chronic health condition.mp.	475
9	chronic health conditions.mp.	2254
10	life limiting conditions.mp.	320
11	life limiting condition.mp.	132
12	(Dynamics of obesity or chronic health conditions).mp.	2285
13	(Multimorbidity or Multimorbidity).mp.	3907
14	6 or 7 or 8 or 9 or 10 or 11 or 12 or 13	276602
15	5 and 14	35219
16	"International Classification of Diseases"/ or classification*.mp. or Classification/ or "International Classification of Functioning, Disability Health"/	95060
17	measure*.mp.	3351266
18	(Directory or Directory).mp.	8965
19	Prevalence/ or prevalence.mp.	685400
20	definition*.mp.	158875
21	identification.mp.	646183
22	identify.mp.	956081

23	linkage.mp.	137523
24	administrative.mp.	56757
25	16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24	5427383
26	15 and 25	8522

5.2.5 Risk of bias in individual studies

A formal critical appraisal of the included studies using a recognised tool (for example, the CASP checklists) was not conducted. This is because the purpose of the review was to describe all available coding systems. A critique of the completeness of the system (for example, the quality of the routine databases on which the data was based, whether the studies included codes for the majority of chronic conditions in childhood, and crucially, whether the codes used were validated in any way) was therefore an integral part of the review process.

5.2.6 Synthesis of results

A narrative synthesis of included studies was conducted to describe the characteristics of the coding systems identified. The review was not examining quantitative data (for example, on the prevalence of or risk factors for a particular chronic condition) and a meta-analysis was therefore not appropriate.

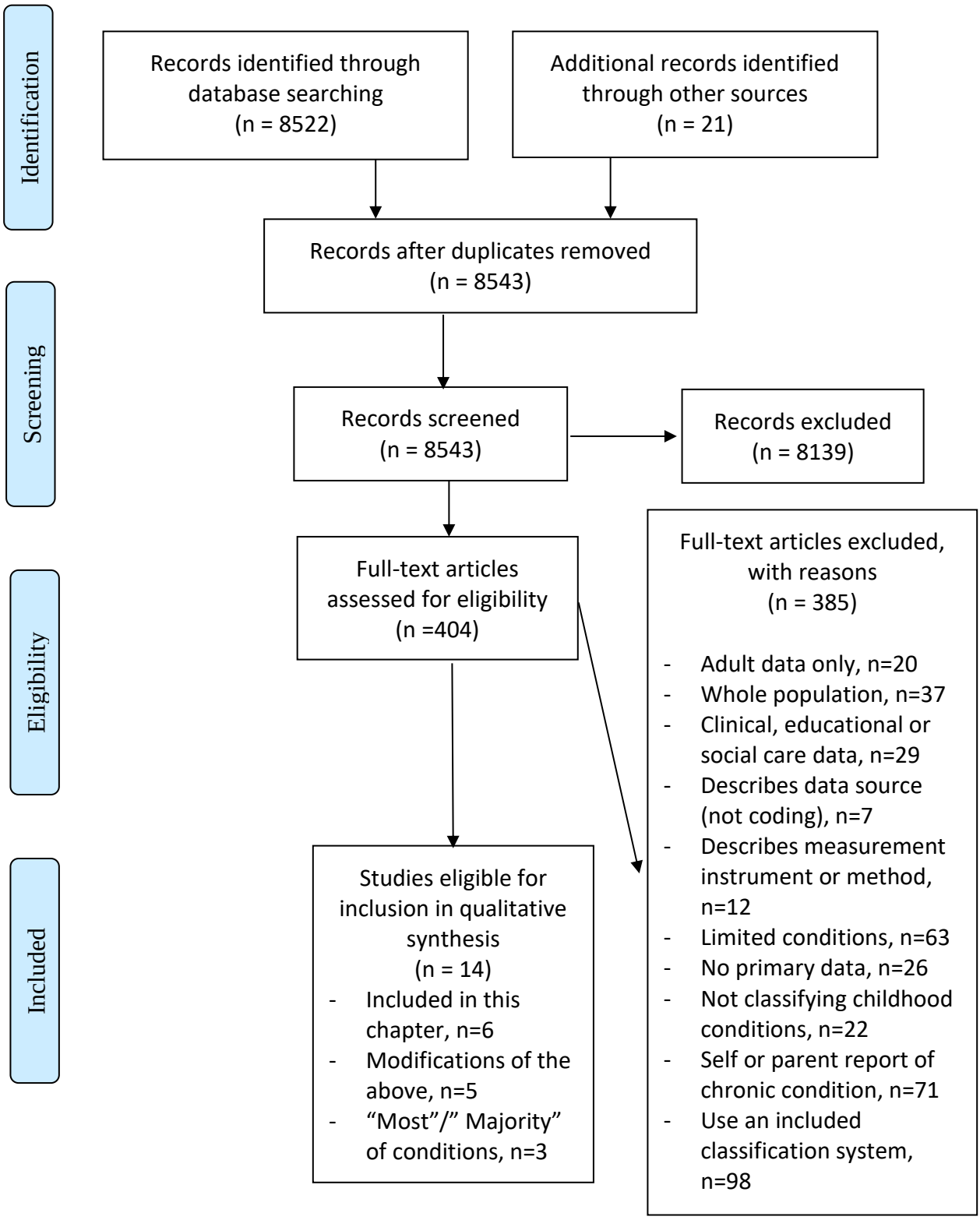
5.3 RESULTS

5.3.1 Study selection

This review summarises the identified existing classification systems for categorizing chronic conditions in children in routine data sources. The terms "chronic health conditions," "chronic diseases and health conditions," and "complex chronic conditions" were found in the included studies. The total number of studies identified through database searching was 8522, with an additional 21 studies identified through other sources. The first screening stage involved reviewing the titles and abstracts of 8543 studies after duplicates were removed. 8139 studies were excluded because they did not meet the inclusion criteria. 404 studies were retrieved for full-text review for eligibility, after which 385 studies were excluded for the following reasons: adult data only (n = 20), whole population (n = 37), clinical, educational, or social care data (n = 29), description of data source but not coding (n = 7), description of measurement instrument or method (n = 12), limited conditions (n = 63), no primary data (n = 26), not classifying

childhood conditions (n = 22), self or parent report of chronic condition (n = 71), and use of an included classification system (n = 98). A total of 14 studies were selected for data extraction and charting: 6 studies attempted to identify all chronic conditions in children in administrative databases, 3 studies attempted to identify most conditions in children, and 5 studies modified other systems by adding in new codes or conditions to identify chronic conditions in children. (See figure 5.1, the PRISMA flow diagram, which summarises the review process).

Figure 5.1: PRISMA flow diagram



5.3.2 Study characteristics

In this review, we summarized the characteristics of the studies that attempted to identify all chronic conditions in children in administrative databases (N = 6) for which data were extracted, including the sample size, PICOS, and their citations (Table 5.2).

Table 5.2: Characteristics for which data were extracted from the included studies

Studies	**Study size**	**PICOS**	**Citations**
Feudtner et al. 2001 Feudtner et al. 2014	Participants included all people 0 to 24 years old in the United States from 1979 to 1997.	A retrospective cohort study using national death certificate data and census estimates from the National Center for Health Statistics.	Feudtner C, Hays RM, Haynes G, Geyer JR, Neff JM, Koepsell TD. Deaths attributed to pediatric complex chronic conditions: national trends and implications for supportive care services. Pediatrics. 2001;107(6). Feudtner C, Feinstein JA, Zhong W, Hall M, Dai D. Pediatric complex chronic conditions classification system version 2: Updated for ICD-10 and complex medical technology dependence and transplantation. BMC Pediatric. 2014;14(1):1–7.
Neff et al. 2002 Neff et al. 2014	34544 children 700 children	A case study was conducted utilising 3M's CRG software and claims data collected and managed by the NWMB, a mid-sized health plan in Washington State, to identify and classify children with a wide range of chronic health conditions. A case study was conducted using CRGs to stratify children receiving care at a tertiary care hospital according to complexity in both hospital and Medicaid administrative data.	Neff JM, Sharp VL, Muldoon J, Graham J, Popalisky J, Gay JC. Identifying and classifying children with chronic conditions using administrative data with the Clinical Risk Group classification system. Ambulatory Pediatric. 2002;2(1):71–9. Neff JM, Clifton H, Popalisky J, Zhou C. Stratification of children by medical complexity. Acad Pediatr [Internet]. 2015;15(2):191–6. Available from: http://dx.doi.org/10.1016/j.acap.2014.10.007
Berry et al. 2017	2.3 million US acute-care	A retrospective cohort analysis of hospital discharges for 0-18 year old children in the Agency for Healthcare	Berry JG, Ash AS, Cohen E, Hasan F, Feudtner C, Hall M. Contributions of Children with Multiple Chronic Conditions to Pediatric Hospitalizations in the United

	hospital discharges	Research and Quality (AHRQ) 2012 Healthcare Cost and Utilization Project Kids' Inpatient Database (KID).	States: A Retrospective Cohort Analysis. Hospital Pediatric. 2017;7(7):365–72. Agency for Healthcare Research and Quality, Healthcare Cost and Utilization project (HCUP). USER GUIDE: CHRONIC CONDITION INDICATOR FOR ICD-10-CM, (BETA VERSION). 2020;1(October).
Hardelid et al. 2014	23 438 children	A retrospective population-based death cohort study using linked death certificates and hospital discharge records for all residents' children in England and Scotland who died between 2001 and 2010, and Wales between 2003 and 2010.	Hardelid P, Dattani N, Gilbert R. Estimating the prevalence of chronic conditions in children who die in England, Scotland and Wales: A data linkage cohort study. BMJ Open. 2014;4(8):1–8.
Mokkink et al. 2008	Not applicable	Systematic review, Theoretical model, and Consensus procedure.	Mokkink LB, Van Der Lee JH, Grootenhuis MA, Offringa M, Heymans HSA. Defining chronic diseases and health conditions in childhood (0-18 years of age): National consensus in the Netherlands. Eur J Pediatric. 2008;167(12):1441–7.
Simon et al. 2014	700 children	A retrospective observational research study was undertaken on children insured by Washington State Medicaid with ≥ 1 emergency department and/or inpatient encounter at Seattle Children's Hospital.	Simon TD, Cawthon ML, Stanford S, Popalisky J, Lyons D, Woodcox P, et al. Pediatric medical complexity algorithm: A new method to stratify children by medical complexity. Pediatrics. 2014;133(6).
Simon et al. 2017	299 children		Simon TD, Cawthon ML, Popalisky J, Mangione-Smith R. Development and Validation of the Pediatric Medical Complexity Algorithm (PMCA) Version 2.0. Hospital Pediatric. 2017;7(7):373–7.
Simon et al. 2018	300 children		Simon TD, Haaland W, Hawley K, Lambka K, Mangione-Smith R. Development and Validation of the Pediatric Medical Complexity Algorithm (PMCA) Version 3.0. Academic Pediatrics [Internet]. 2018;18(5):577–80. Available from: https://doi.org/10.1016/j.acap.2018.02.010

5.3.3 Results of individual studies

In this review, we presented the data sources used in each study, including the type data sources, and classification systems (e.g., ICD-9, ICD-10, etc.) of the chronic conditions in children and provided links to their associated codes (if available) in Table 5.3.

Table 5.3: A summary of the chronic conditions in children according to each study

Studies	**Data sources**	**Type of Classification systems**	**Type of Codes**	**Codes**
Feudtner et al. 2001 Feudtner et al. 2014	National death certificate data and census estimates from the National Centre for Health Statistics.	The paediatric complex chronic conditions (CCC) classification system	ICD-9-CM ICD-10	https://static-content.springer.com/esm/art%3A10.1186%2F1471-2431-14-199/MediaObjects/12887_2014_1131_MOESM5_ESM.docx
Neff et al. 2002 Neff et al. 2014	Claims data	The Clinical Risk Groups (CRGs) classification system	ICD-9-CM ICD-9	It is unclear whether the codes are publicly accessible.
Berry et al. 2017	Administrative data: insurance claims data	Agency for Healthcare Research and Quality Chronic Condition Indicator (CCI) (beta version)	ICD-10-CM	https://www.hcup-us.ahrq.gov/toolssoftware/chronic_icd10/CCI-ICD10CM-v2021-1.zip
Hardelid et al. 2014	Linked death certificates and hospital discharge records	The Hardelid et al. classification	ICD-10	bmjopen-2014-005331supp_tables.pdf
Mokkink et al.	Hospital data	Mokkink et al. classification	ICD-10	https://static-content.springer.com/esm/art%3A1

2008				0.1007%2Fs00431-008-0697-y/MediaObjects/431_2008_697_MOESM1_ESM.doc
Simon et al. 2014 Simon et al. 2017 Simon et al. 2018	Hospital discharge data and Medicaid claims data	The Pediatric Medical Complexity Algorithm	ICD-9-CM ICD9/10-CM	The codes found in this link: https://github.com/kpwhri/pmca/raw/main/docs/pmca_dx_code_lists.xlsx

The six classification systems identified from the studies (N = 6) that attempted to identify all chronic conditions in children are: 1) the paediatric complex chronic conditions (CCCs) classification system, 2) the clinical risk group (CRGs) classification system, 3) the Chronic Condition Indicator (CCI) developed by the Agency for Healthcare Research and Quality (AHRQ), 4) the Hardelid et al. classification of chronic conditions in children, 5) the national definition of chronic diseases and health conditions in the Netherlands, and 6) the Pediatric Medical Complexity Algorithm (PMCA). I presented below a background on the development of these 6 classification systems, the type of routine data sources and other administrative dataset used, the coding systems upon which they are based (for example, ICD-10), and the groupings used (for example, chapter headings within ICD-10), and, if possible, the specific codes used to identify a specific condition.

5.3.4 Background on the development of the six classification systems

1) The paediatric complex chronic conditions (CCCs) classification system:

The paediatric complex chronic conditions (CCCs) classification system version 1 (v1) was created by Feudtner et al. to identify how deaths attributable to paediatric. CCCs have changed over the last two decades of 2000, to examine the numbers and rates of CCCs-attributed deaths by cause and age at the time of death, and to determine the average number of children living within the last 6 months of their lives [179].

2) The clinical risk group (CRGs) classification system:

Neff et al. utilised available 3M's Clinical Risk Groups (CRGs) software to stratify children, in a tertiary children's hospital, Seattle Children's Hospital (SCH), and a state's Medicaid claims data, Washington State (WSM), into 3 condition groups: complex chronic disease (C-CD);

noncomplex chronic disease (NC-CD), and nonchronic disease (NC) [180]. The target gold standard population of 700 children was created from a random sample of 1000 children, which was divided into three groups: 350 C-CD, 100 NC-CD, and 250 NC, using electronic medical records and expert consensus. WSM data consisted of encounters and claims for all care sites. The SCH data only contained inpatient, emergency department, and day surgery claims [180].

3) The Chronic Condition Indicator (CCI) developed by the Agency for Healthcare Research and Quality (AHRQ):

The first version of the Chronic Condition Indicator (CCI) and the Clinical Classification System (CCS), for ICD-9-CM was developed as part of the Agency for Healthcare Research and Quality's (AHRQ) Healthcare Cost and Utilization Project (HCUP) to support health services research by allowing researchers to quickly detect diagnoses that indicate a chronic condition [181]. AHRQ modified the first version of the CCI (i.e., ICD-9-CM) to ICD-10-CM (beta version), and updated it annually to correspond with fiscal year revisions to the ICD-10-CM coding system. The beta version remained consistent with the ICD-9-CM version's definition of a chronic condition [182]. Berry et al. adapted the AHRQ's open-source, publicly available diagnosis classification system to improve its utility in children as the following:

"1. creating 8 new organ systems (e.g., metabolism) that were originally grouped with a more heterogeneous organ system (e.g., genetics and/or metabolism), leading to a total of 25 organ systems; 2. distinguishing 159 new distinct CCS categories of chronic conditions (e.g., gastroesophageal reflux disease) by ICD-9-CM codes that were grouped non-specifically (e.g., other digestive disorders) into an expanded list of more specific categories, leading to a total of 531 categories of chronic conditions (372 existing and 149 new); 3. redesignating 39 CCS diagnoses from chronic to nonchronic when applied to children (e.g., cystitis); and 4. reassigning 13 CCS categories from nonchronic to chronic (e.g., chronic kidney disease)" [183].

4) The Hardelid et al. classification of chronic conditions in children:

Hardelid et al. developed a definition of chronic conditions in children who died in order to use it with longitudinal hospital discharge data linked to death certificates to estimate the proportion of children aged 1–18 years who died with chronic conditions in England, Scotland, and Wales and to investigate time trends in childhood deaths involving chronic conditions [178].

5) The national definition of chronic diseases and health conditions in the Netherlands:

A national consensus in the Netherlands has been obtained on a comprehensive definition of chronic conditions in children, which could be used as a framework for epidemiological studies. The consensus was based on a thorough review of the literature on definitions of chronic conditions in childhood and a theoretical framework of determinants and indicators of a population's health status. The definition was subsequently revised based on feedback from 21 Dutch experts, including clinicians, researchers, and representatives of patient organisations, until agreement was reached [171].

6) The Pediatric Medical Complexity Algorithm (PMCA):

From 2014 to 2018, Simon et al. developed three versions of the Pediatric Medical Complexity Algorithm (PMCA) [184] [185] [186]. Simon et al. developed PMCA version 3.0 by modifying PMCA version 2.0 to include the International Classification of Diseases Ninth and Tenth Revisions and Clinical Modification (ICD9/10-CM) codes for classifying children with chronic disease (CD) according to their level of medical complexity [37]. They applied PMCA version 3.0 to Seattle Children's Hospital data for children with ≥1 emergency department (ED), day surgery, and/or inpatient encounter from January 1, 2016, to June 30, 2017. Starting with the encounter date, up to 3 years of retrospective discharge data were used to classify children as having complex chronic disease (CCD), noncomplex chronic disease (NC-CD), and no CD [186].

5.3.5 The definitions and categories of chronic conditions in children

There were some differences and some similarities across all the definitions of chronic conditions in children used in the six classification systems. The majority of the definitions specified the time frame for which a chronic condition is expected to last (e.g., at least 12 months [187] [184]), be present (e.g., for longer than 3 months [171]), or require follow-up (e.g., more than 1 year [178]). All the six definitions did not agree on the number of organ systems that should be included (i.e., whether one or several organ systems). Also, they did not agree on the age of children (e.g., only the national definition of chronic diseases and health conditions in the Netherlands specified the age of children to be 0-18 years. In addition, each classification system has its own unique categories of chronic conditions in children (See table 5.4).

Table 5.4: The definitions and categories of chronic conditions in children used in the six classification systems

The classification system	**Definitions**	**The categories of the chronic conditions in children**
The paediatric complex chronic conditions (CCCs) classification system	Feudtner et al. defined complex chronic conditions (CCCs) in children as *"any medical condition that can be reasonably expected to last at least 12 months (unless death intervenes) and to involve either several different organ systems or one organ system severely enough to require specialty paediatric care and probably some period of hospitalization in a tertiary care centre"* [187].	Feudtner et al. classified CCCs into 9 categories (cardiovascular, respiratory, neuromuscular, renal, gastrointestinal, hematologic or immunologic, metabolic, other congenital or genetic, and malignancy) [187].
The clinical risk group (CRGs) classification system	Neff et al. used the same consensus definition of conditions as detailed in a separate COE4CCN-funded study of a publicly available software, Pediatric Medical Complexity Algorithm (PMCA) [180].	
The Chronic Condition Indicator (CCI) developed by the Agency for Healthcare Research and Quality (AHRQ)	Berry et al. defined multiple chronic conditions as *" 2 or more chronic conditions that affect 2 or more organ systems"* [183]. Initially, the CCI classified conditions as chronic or non-chronic, then it was expanded to identify four different types of conditions beginning with v2021.1 (beta version) [182]: 1) Chronic: Examples include malignant cancer, diabetes, obesity, hypertension, and many mental health conditions. 2) Acute: Examples include aortic embolism, bacterial infection, pregnancy, and an initial encounter for an injury. 3) Both: Examples include persistent asthma with (acute) exacerbation, acute on chronic heart failure, and kidney transplant rejection. 4) Not applicable (code cannot be used to identify a chronic or	In 2002, Neff et al. used the 3M's CRG software to stratify the children by severity of chronic conditions into 9 hierarchical, mutually exclusive health status groups: 1) nonchronic acute; 2) nonchronic significant acute; 3) minor chronic; 4) multiple minor chronic; 5) moderate chronic or dominant chronic in only a single body system; 6) moderate chronic or dominant chronic in 2 body systems; 7) dominant chronic in 3 or more body systems; 8) malignancies receiving active therapy; and 9) catastrophic conditions that are progressive, receipt of solid organ transplant, or long-term dependency on technology [188]. However, in 2014, Neff et al. used the 3M's CRG v1.9 software to stratify children according to complexity into three categories: complex chronic disease (C-CD), noncomplex chronic disease (NC-CD), and nonchronic disease (NC) [182].

	acute condition): Examples include external cause of morbidity codes, injury sequela codes, and codes starting with the letter Z for screening or observation.	
The Hardelid et al. classification of chronic conditions in children	Hardelid et al. defined a chronic condition as *"any health problem likely to require follow-up for more than 1 year, where follow-up could be repeated hospital admission, specialist follow-up through outpatient department visits, medication, or use of support services"* [178].	
The national definition of chronic diseases and health conditions in the Netherlands	The Netherlands has agreed on a comprehensive definition of chronic conditions in children based on four criteria: *"a disease or condition is considered to be a chronic condition in childhood if: (1) it occurs in children aged 0 up to 18 years; (2) the diagnosis is based on medical scientific knowledge and can be established using reproducible and valid methods or instruments according to professional standards; (3) it is not (yet) curable or, for mental health conditions, if it is highly resistant to treatment and (4) it has been present for longer than three months or it will, very probably, last longer than three months, or it has occurred three times or more during the past year and will probably reoccur"* [171].	Berry et al. presented the following 20 (out of 531) most common chronic conditions among US paediatric hospitalizations in 2012, which were coded during US hospital discharges for children (2.3 million) and classified by the adapted AHRQ CCI and CCS for children: asthma, substance-abuse disorder, depression, oesophageal reflux, epilepsy, anxiety disorder, ADHD, bipolar disorder, enterostomy, obesity, type 1 diabetes, sickle cell anaemia, hypertension, cardiac dysrhythmias, cerebral palsy, oppositional defiant disorder, pervasive developmental disorder, schizophrenia, developmental disabilities, and hypothyroidism [13].
The Pediatric Medical Complexity Algorithm (PMCA)	The PMCA was based on the Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN) consensus definitions for three levels of medical complexity [184]: 1) Children with C-CD: *"Significant chronic conditions in two or more body systems. A significant chronic condition is defined as a physical, mental, or developmental condition that can be expected to last at least a year, will use health care resources above the level for a healthy child, require treatment of control of the condition, and the condition can be expected to be episodically or continuously debilitating,*	

	or a progressive condition that is associated with deteriorating health with a decreased life expectancy in adulthood, or continuous dependence on technology for at least 6 months, or malignancies: progressive or metastatic malignancies that affect life function. Exclude those in remission for >5 years". 2) Children with NC-CD: "*Chronic conditions that last at least 1 year. These conditions are commonly lifelong but can be episodic with periods of good health between episodes. They include physical, developmental, or mental health conditions that may persist into adulthood but may also resolve either secondary to the natural history of the disease or as a result of surgical intervention. These conditions involve a single body system, are not progressive, can vary widely in severity, and result in highly variable health care utilization".* 3) Children without CD: *"1) Acute nonchronic conditions: a physical, developmental or mental health condition that is not expected to last >1 year. These children may temporarily (for <1 year) use health care resources above the normal level for a healthy child, and 2) Healthy: No acute or chronic health conditions. These children do not use health care resources above the normal level for a healthy child".*	

All the six classification systems were based on the International Classification of Disease (ICD) coding system including ICD-9, ICD-9-CM, ICD-10, and ICD-10-CM. They all used administrative data sources to identify cases of children with chronic conditions including national death certificate data, census estimates, claims data, and hospital discharge records (See table 5.3).

5.3.6 Estimates of the prevalence of chronic conditions in children

According to Mokkink et al., prevalence estimates for chronic health conditions in children in the literature ranged from 0.22% to 44%, depending on the definitions and operationalizations used [166]. This supports our findings from the six papers. There were six different ways of classifying chronic conditions in children. This meant that estimates of prevalence differed

greatly from 29.3 % of those classified as having CMCC in Berry et al. [183] to 34% classified as having complex chronic disease in Simon et al. [186]; and from to 33% classified as without chronic disease in Simon et al. [186] to 34% classified as having no chronic conditions in Berry et al [183] to 85% classified as healthy in Neff et al. [188].

5.3.7 The six classification systems' strategies for dealing with multiple conditions in the same child

Feudtner et al. sated that the CCCs classification system is flexible in that it can be used to investigate a single CCCs category or to identify patients with several CCCs categories and significant multisystem comorbidities [189]. Neff et al. (2002) mentioned that each child is counted only once within each condition group, but may be present in multiple condition groups [188]. Berry et al. generated for each hospitalised child a chronic condition profile (of 51 total) grouped by the organ system allocated to each chronic condition (varying from none to multiple) was identified [183]. Hardelid et al. reported that children with multiple conditions were included in each group for condition-specific analysis, but were counted only once for overall analyses [178]. Mokkink et al. and Simon et al. did not describe how they count multiple chronic conditions in the same child based on their definitions; therefore, it is unclear how this affects the prevalence data for each cause [171] [186].

5.3.8 Data on further sub-divisions of chronic conditions or by causes

Most classification systems gave data on further sub-divisions of chronic conditions in children. Feudtner et al. discovered that the proportion of all deaths attributed to each cause varied by age: 0.25% of infant deaths, 20% of childhood deaths, and 7% of adolescent deaths were caused by noncancer CCCs; cancer CCCs were responsible for 1% of infant deaths, 11% of childhood deaths, and 6% of adolescent deaths; and injuries accounted for 3% of infant deaths, 47% of childhood deaths, and 76% of adolescent deaths [179]. Also, Hardelid et al. reported the proportion of children who died with each type of chronic condition according to the amount of data used from death certificates and longitudinal hospital records by age group [178]:

1) Cancer and blood disorders were the most prevalent chronic conditions recorded based only on the underlying cause of death (17.6% of child deaths aged 1–18 years; 4123 out of 23 438 children).
2) Mental/behavioral conditions were the most prevalent chronic conditions in children aged 15–18. Using death certificates and hospital data up to a year

before death, 26.5 5% of children in this age group died with a mental/behavioural disorder (2557 out of 9663).

3) Adding data from hospital records up to one year before death resulted in a two (in 10-14-year-olds) to eight (in 1-4-year-olds) time increase in the proportion of children who died with mental/behavioral disorders among children younger than 15 years.
4) Only 31–49 % of children with chronic infections, respiratory, metabolic/endocrine/renal/digestive/GU, musculoskeletal/skin, and cardiac conditions would have been detected using death certificates, compared to also using hospital records up to 1 year before death. Only 31–49% of children with chronic infections, respiratory, metabolic/endocrine/renal/digestive/GU, musculoskeletal/skin, and cardiac conditions would have been detected using death certificates, compared to also using hospital records up to 1 year before death.

Berry et al. investigated the relationship between patient demographic characteristics and the prevalence of chronic conditions in hospitalised children. These were admission age in years, type of insurance (public, private, and more), and race and ethnicity (Asian American and American Indian, Hispanic, non-Hispanic African American, non-Hispanic white, and more). They discovered that the median age of children hospitalised without a chronic condition (3 years) was lower than the ages of children hospitalised with one chronic condition (10 years) and multiple chronic conditions (12 years) [183].

Neff et al. (2002) presented the prevalence rates for selected chronic condition groups, including asthma (1.97 %), attention deficit hyperactivity disorder (3.08 %), cystic fibrosis (0.03 %), cerebral palsy (0.02 %), diabetes (0.17 %), learning disorders (0.39 %), malignancies (0.07 %), mental health conditions (5.54 %), and mental retardation (0.25 %) [188]. According to Simon et al., conditions such as cystic fibrosis, complex congenital heart disease, and malignancy were flagged as progressive. Body system flags were also assigned to enable body system counts and subsequent classification as NC-CD (1 body system) or C-CD (2 body systems) [186]

5.3.9 The six classification systems validations of the identified cases of children with chronic conditions

1. ***The paediatric complex chronic conditions (CCCs) classification system:***

   Feudtner et al. used the CCCs version 2 (v2) system to classify all cases in the CDC Multiple Cause of Death data for 1996 (http://wonder.cdc.gov/mortSQL.html), 2009 Kids' Inpatient Database (KID), and 2010 Nationwide Emergency Department Sample (NEDS) datasets, and analysed all cases classified as not having a CCC to assess whether any codes in the CCC v2 system were either incorrectly specified or omitted, and then they fixed any errors. Also, they assessed the CCCs v2 system's performance by comparing its performance across ICD-9 and ICD-10 codes and comparing its performance year-to-year before and after ICD-10 implementation [189].

2. ***The clinical risk group (CRGs) classification system:***

   The CRGs classification results were validated by Neff et al. (2002), who compared the prevalence rates for specific conditions estimated using CRGs to those published in the general paediatric literature. Asthma, ADHD, learning difficulties, mental health conditions, and mental retardation were selected according to their relative frequency in children, while cystic fibrosis, cerebral palsy, and diabetes were selected due to their status as sentinel chronic conditions that are resource-intensive and life-long. Neff et al. (2014) reported that CRGs worked well in a three-way stratification of hospitalised children, with the exception of low sensitivity for identifying the NC-CD group, and it demonstrated high sensitivity and specificity in identifying C-CD. They mentioned that the low sensitivity for identifying NC-CD may be explained in part by the imprecision of ICD-9 coding for emergency department and day surgery data, as well as the difficulty of translating into ICD-9 methodology whether conditions may or may not become chronic [180].

3. ***The Chronic Condition Indicator (CCI) developed by the Agency for Healthcare Research and Quality (AHRQ):***

   AHRQ system, was independently reviewed by two complex-care paediatricians with experience caring for children with a wide range of chronic conditions. They evaluated the face validity of the organ system categories first, then the CCS diagnosis categories, and finally their related ICD-9-CM codes as they apply to children. The paediatricians then examined ICD-9-CM codes that were not classified as chronic conditions, and they suggested improvements that were examined by a wider study committee that also

adjudicated the few cases of disagreement between these two independent assessments [183].

4. ***The Hardelid et al. classification of chronic conditions in children:***

   Hardelid et al. validated their definition of chronic conditions by comparing the prevalence of chronic conditions in children who died to the prevalence of chronic conditions in all children admitted to hospital, and they recommended conducting additional research to test their definition in other settings, such as intensive care and primary care databases [178].

5. ***The national definition of chronic diseases and health conditions in the Netherlands:***

   Mokkink et al. reported that three paediatricians assessed all ICD-10 diagnoses informally. Except for diagnoses classified as (other) or (unspecified), *"each diagnosis was assessed using all four criteria of the consensus definition to determine whether it denoted (+) a chronic condition occurring in childhood, (−) a condition that is either not chronic or does not occur in childhood, or (+) a condition that is chronic in just a proportion of the affected patients"* [171].

6. ***The Pediatric Medical Complexity Algorithm (PMCA):***

   Simon et al. performed validity testing by determining the sensitivity and specificity of all three versions of PMCA for correctly classifying children (who had accessed tertiary hospital care) into the three levels of complexity using Seattle Children's Hospital discharge and Medicaid claims data and comparing these categorizations to those obtained through medical record review (the gold standard). They concluded that all three versions of PMCA identified children with C-CD with good sensitivity and good to excellent specificity[184] [185] [186].

5.4 Discussion

5.4.1 Summary of evidence:

The majority of studies (N = 6) included in this review, which attempted to identify all chronic conditions in children, were conducted in the United States. We believe that the abundance of studies from the United States is due to the administrative data they have such as the insurance datasets, which need to have a diagnosis recorded (usually based on ICD coding systems) in order to process claims appropriately [190]. Although these insurance datasets provide a record of chronic conditions, many of the USA studies linked the insurance datasets to hospital-based datasets to capture the target cases [178] [184] [185] [186].

The data sources in the United States are slightly different than those in the United Kingdom, as their health care system lacks insurance datasets. As a result, researchers cannot identify all chronic conditions in children simply by relying on hospital data based on ICD coding systems, as not all children with chronic conditions would be admitted to hospitals. Thus, in the UK, ICD coding systems are beneficial only for hospital records but not for primary care records, which researchers require if they are to examine all chronic conditions in children. Accordingly, several data linkages between hospital records and GP records have been developed. While the data sources differ between the USA and the UK, researchers in the UK may be interested in learning about the available classification systems for chronic conditions in children from the studies that might be useful for some circumstances, which clarify why they were included in this review.

5.4.2 Specific discussion points:

I found that not all the studies provided an open link to the list of codes in their manuscripts, including: 1) the Clinical Risk Groups (CRGs) classification system, which may be because 3M CRGs has now been developed into a commercial product [180]; 2) the PMCA, which we wouldn't be able to find easily or straightforwardly (e.g., we searched by author names on different websites to locate the codes of the PMCA [191], which we located at this link: https://www.kpwashingtonresearch.org/our-research/our-scientists/rita-mangione-smith-md-mph/measurement-tools-research-dr-rita-mangione-smith; 3) and the paediatric complex chronic conditions (CCCs) classification system provided a non-open link in the related manuscript to the lists of codes [35]. This illustrates the difficulties that researchers have when attempting to locate some of the published algorithms with their codes. This may appear

simple, but when researchers go through the search procedure, they find it quite difficult (i.e., time-and labour-consuming), and they occasionally find algorithms that are not suitable for their studies (e.g., they are not reusable because they are based on different data sources). Complex algorithms that combine multiple data sources to identify every single case of a disease are probably the best option, but they are extremely time- and labour-intensive. For example, Al Sallakh did a doctoral project in Wales that used routinely collected EHR data to identify and classify patients who had asthma in Wales. He found that the absence of a consensus on how to use EHR data to define asthma or measure asthma outcomes and that the different types of asthma and the limitations of routinely collected EHR data make it difficult to use these data for this purpose. In addition, He conducted a systematic search for asthma-related literature published between January 1, 2014 and December 31, 2015, and he extracted the algorithms used to identify asthma patients and assess severity, control, and exacerbations from the eligible studies using various forms of EHR-derived data sources. He discovered high variation in the algorithms used to define asthma (n = 66 different algorithms) across 113 eligible articles. [191].

The analysis of the results demonstrated that there is no agreed-upon definition of "chronic conditions" in children globally, and some of the most widely used classification systems in the United States, such as the paediatric complex chronic conditions (CCCs) classification system [189], are not completely inclusive. For example, CCCs do not include asthma, obesity, attention deficit hyperactivity disorder, or other behavioral health conditions (e.g., depression, bipolar disorder) because they focus on either “medical complexity” or “children with special healthcare needs” rather than all children with chronic conditions [192]. As a result, several modifications have to be made by researchers to incorporate important conditions such as obesity and mental health conditions. Similarly, the researchers in the UK are unsure how inclusive they should be in defining chronic conditions in children. They are unsure whether they should include or exclude (e.g., atopy, eczema, transplant patients, cancer survivors, adverse pregnancy outcomes, preterm birth, and obesity; if they include obesity, then they should also include malnutrition). As a result, we excluded in our study other classification systems reported in the literature that examined a specified set of chronic conditions in children (e.g., children with life-limiting conditions or, children with disabilities) [193] [194] [195] [196].

Based on the findings of this review, we suggest that the UK needs to reach a national consensus on: 1) what constitutes a "chronic condition" in children, 2) the purpose of a definition is (e.g., some of these conditions require regular healthcare input while others do not), and 3) the parameters for the condition are (e.g., a recognisable diagnosable condition or something that increases a child's risk of having poor health outcomes in the future?). A national consensus on these areas could help to maximise the use of routine data sources, assure consistency across studies, and facilitate statistical data analysis across the healthcare system [33]. Relating to this, the Royal College of Paediatrics and Child Health and CHR-UK Programme of Work at the MRC Centre of Epidemiology for Child Health, University College London Institute of Child Health developed a definition of chronic conditions in children, who die in routine mortality and hospital admission data [197], which could be adapted by the NHS.

The use of a consistent classification system to identify chronic conditions in children would be ideal, but this is impracticable due to the wide variety of data sources. The USA studies, in particular, rely on insurance databases that require the identification of specific conditions, which makes it reasonably easy to identify children with chronic conditions (assuming that the reporting of these conditions is correct and complete) [40].

This is not the case in a setting such as the UK, where routine data do not always reveal which conditions a child may have. For example, hospital-based data, which provides up to 16 codes, indicates that chronic conditions should be clear but are not always. Hardelid et al. stated that *"ICD-10 codes alone were not sufficient to determine whether certain conditions, particularly injuries and cardiac conditions, were acute or could lead to chronic sequelae"* [197]. Furthermore, data for conditions where children are not frequently hospitalised, such as neurodevelopmental conditions, is incomplete compared to GP or outpatient data, depending on where they are managed, which is based on random coding.

5.4.3 Strengths and limitations

Conducting a thorough literature search on such a broad topic as chronic conditions in children was challenging, and some potentially eligible studies were excluded because they did not publish code lists. These lists must be released publicly in order for studies to be reproduced or differences in results to be acknowledged. For example, researchers who combined hospital diagnoses using ICD 10 with medications from GP records would get different lists of codes than researchers who combined them with diagnosis codes (i.e., Read codes) within GP records.

This review was based on searching only one database (Medline). We selected this database because it has the best coverage for health-related studies, which is where we assumed most studies in our area of interest would be indexed. Despite only using one database, our search identified more than 8000 records, and we screened more than 400 full-text studies. Many of these used one of the six classification systems described in the review. We also checked the reference lists of all studies that had used a seemingly relevant classification system, to ensure that we assessed any source studies against our eligibility criteria. Therefore, whilst it is possible that we would have identified other relevant coding systems if we had repeated the searches in other databases, we are reasonably confident that our search has identified most published papers that describe such systems. However, some judgement was required in selecting individual studies due to the considerable differences in terminology, study types, and study designs found in the literature.

We feel that this review will be useful to researchers interested in identifying chronic conditions in children from routine data sources such as administrative databases. A summary of some of the available classification systems for chronic conditions in children, including links to code lists, would save researchers time and effort. For example, researchers may use the algorithms that use an ICD-based classification system to identify chronic conditions in data on hospitalisations or deaths in the UK, but only if the system matches or can be edited to fit their purposes. This review may also be useful for researchers interested in examining studies that classify a range of chronic conditions in children using a range of different combined data sources. However, additional work is needed to make sure that current systems that try to map ICD-10 codes to other systems for chronic conditions in children are validated.

This review may be beneficial for concept libraries such as clinicalcode.org and CALIBER [143] [18] as we tried to make the classification systems presented in this review repeatable by summarising the definitions, types of routine data sources, coding systems on which they are based (for example, ICD-10), and links to their specific codes if they are publicly available.

5.5 Conclusion

The wide range of terms, definitions, and classification systems for chronic conditions in children that we found demonstrates that there is no consensus among researchers in the majority of studies on these conditions. In reality, no single standard classification system is

likely to be able to serve all of the different research questions. Instead, an ideal system should be flexible and capable of being modified to address each specific research question.

The classification systems, in order to be useful for identifying chronic conditions in children, should also be transparent in terms of how they are constructed, implemented, and afterwards audited, which can be facilitated by providing open-source code that can be reviewed and modified by end-users. Researchers would examine classification systems developed in the United States to determine whether they would be applicable and useful in the UK (e.g., Wales). In conclusion, it is recommended that researchers in the UK collaborate to develop an algorithm that identifies all chronic conditions in children from various data sources (e.g., hospital data, GP records, or GP pharmaceutical records) and a framework that contains a hierarchy of all chronic conditions in children to be considered as a gold standard that links each condition with all possible data sources.

Table 5.5: Criteria for eligibility, study characteristics, data sources, and definitions to identify studies of chronic conditions in children

Table 5.5.a: Eligibility

First author & year published	**Journal**	**Publication type**	**Inclusion criteria 1: Discuss MEASUREMENT of chronic conditions in childhood**	**Inclusion criteria 2: Include information on MORE THAN ONE chronic condition**	**Inclusion criteria 3: Provide details of how the chronic conditions were CLASSIFIED/DEFINED**	**Eligible for inclusion**	**Notes**	**Reviewer initials**

Table 5.5.b: Study Characteristics

First author & year published	Country	Geographic region	Data collection setting	Sample size	Description of participants	Study design	Study aim (copy verbatim from paper)	Study time period	Age of participants: (Mean [SD] or range)	Gender: Female (N, %)	Reviewer initials

Table 5.5.c: Data Sources

First author & year published	Data on chronic conditions from DISEASE REGISTER?	Data on chronic conditions from HOSPITAL ADMISSIONS?	Data on chronic conditions from PRIMARY CARE VISITS?	Data on chronic conditions from DEATH REGISTRATION?	Data on chronic conditions from HEALTH INSURANCE RECORDS?	OTHER source of data on chronic conditions [Please specify]	Data source used ICD-10 codes to identify chronic conditions?	Data source used ICD-9 codes to identify chronic conditions?	Data source used READ codes to identify chronic conditions?	Data source used ANOTHER SYSTEM to classify chronic conditions [Please specify]	Classification system used to group chronic conditions [e.g., ICD-10-chapter headings, or classification system devised by authors]	Any clinical input in defining or grouping the chronic conditions?	If yes to question M, please specify [copy verbatim from paper if possible]

Table 5.5.d: Definitions

Link to ICD-10	**Infections and parasitic diseases (chronic only; ICD-10 chapters A & B)**			**Neoplasms (ICD-10 chapters C & D)**		
First author & year published	**Included**	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]	**Included**	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]

Table 5.5.d: Definitions (continued)

Diseases of the blood and blood-forming organs (ICD-10 D50-D89)			**Endocrine, nutritional and metabolic diseases (ICD-10-chapter E)**			**Mental and behavioural disorders (ICD-10-chapter F)**		
Included	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]	**Included**	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]	**Included**	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]

Table 5.5.d: Definitions (continued)

Diseases of the nervous system (ICD-10-chapter G)			**Diseases of eye and adnexa, or ear (ICD-10-chapter H)**			**Diseases of the circulatory system (ICD-10 chapter I)**		
Included	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]	**Included**	**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]		**Chronic conditions included** [Only list individual conditions if an accepted system such as ICD-10 is NOT used]	**Definition** [Only list individual definitions if an accepted system such as ICD-10 is NOT used]

Chapter 6

Results 1 - Qualitative study

This chapter presents the results of the two qualitative studies that aimed to explore the requirements for a portal of disease phenotyping definitions (a concept library) for users from a variety of disciplines, including researchers, clinicians, machine learning experts, and research managers. This chapter is based on the following published paper in the JMIR Human Factors on 15/3/2022**:** Concept Libraries for Repeatable and Reusable Research: Qualitative Study Exploring the Needs of Users, which can be accessed via http://dx.doi.org/10.2196/31021

Authorship Declaration	
Candidate	**Name and College**
Author 1	Zahra Almowil, Data Science Building, Medical School, Swansea University, Wales SA2 8PP
Author 2	Shang-Ming Zhou, Centre for Health Technology, Faculty of Health, University of Plymouth, Plymouth, PL4 8AA, UK
Author 3	Sinead Brophy, Data Science Building, Medical School, Swansea University, Wales SA2 8PP
Author 4	Jodie Croxall, Data Science Building, Medical School, Swansea University, Wales SA2 8PP

Author Details and their Roles:

Paper 2 (Concept Libraries for Repeatable and Reusable Research: Qualitative Study Exploring the Needs of Users)

Located in Chapter 6

The candidate conceived and designed the research, collected and analysed the data, and wrote the first draft and re-drafts of the paper. Her work was supervised by Shang-Ming Zhou, Sinead Brophy, and Jodie Croxall. We also independently read and agreed on the data extract. Shang-Ming Zhou, Sinead Brophy, and Jodie Croxall advised on the analysis. The work in this paper was 85% that of the candidate and 15% that of others.

We the undersigned agree with the above stated “proportion of work undertaken” for each of the above published peer-reviewed manuscripts contributing to this thesis:

Signed Candidate

Author 1: Zahra Almowil
Author 2: Shang-Ming Zhou
Author 3: Sinead Brophy
Author 4: Jodie Croxall

Original Paper

Concept Libraries for Repeatable and Reusable Research: Qualitative Study Exploring the Needs of Users

Zahra Almowil[1], MSc; Shang-Ming Zhou[2], PhD; Sinead Brophy[1], PhD; Jodie Croxall[1], PhD

[1]Data Science Building, Medical School, Swansea University, Swansea, Wales, United Kingdom

[2]Centre For Health Technology, Faculty of Health, University of Plymouth, Plymouth, United Kingdom

Corresponding Author:
Zahra Almowil, MSc
Data Science Building
Medical School, Swansea University
Sketty
Swansea, Wales, SA2 8PP
United Kingdom
Phone: 44 07552894384
Email: 934467@swansea.ac.uk

Abstract

Background: Big data research in the field of health sciences is hindered by a lack of agreement on how to identify and define different conditions and their medications. This means that researchers and health professionals often have different phenotype definitions for the same condition. This lack of agreement makes it difficult to compare different study findings and hinders the ability to conduct repeatable and reusable research.

Objective: This study aims to examine the requirements of various users, such as researchers, clinicians, machine learning experts, and managers, in the development of a data portal for phenotypes (a concept library).

Methods: This was a qualitative study using interviews and focus group discussion. One-to-one interviews were conducted with researchers, clinicians, machine learning experts, and senior research managers in health data science (N=6) to explore their specific needs in the development of a concept library. In addition, a focus group discussion with researchers (N=14) working with the Secured Anonymized Information Linkage databank, a national eHealth data linkage infrastructure, was held to perform a SWOT (strengths, weaknesses, opportunities, and threats) analysis for the phenotyping system and the proposed concept library. The interviews and focus group discussion were transcribed verbatim, and 2 thematic analyses were performed.

Results: Most of the participants thought that the prototype concept library would be a very helpful resource for conducting repeatable research, but they specified that many requirements are needed before its development. Although all the participants stated that they were aware of some existing concept libraries, most of them expressed negative perceptions about them. The participants mentioned several facilitators that would stimulate them to share their work and reuse the work of others, and they pointed out several barriers that could inhibit them from sharing their work and reusing the work of others. The participants suggested some developments that they would like to see to improve reproducible research output using routine data.

Conclusions: The study indicated that most interviewees valued a concept library for phenotypes. However, only half of the participants felt that they would contribute by providing definitions for the concept library, and they reported many barriers regarding sharing their work on a publicly accessible platform. Analysis of interviews and the focus group discussion revealed that different stakeholders have different requirements, facilitators, barriers, and concerns about a prototype concept library.

(JMIR Hum Factors 2022;9(1):e31021) doi: 10.2196/31021

KEYWORDS

electronic health records; record linkage; reproducible research; clinical codes; concept libraries

Introduction

Background

Health care systems are becoming more digitally focused rather than paper-based and are moving to the use of electronic health records (EHRs) [1]. This means there is a large amount of electronic patient data that can be moved and linked together into safe data repositories to enable researchers and data analysts to query and examine these data effectively [2-5]. The growing availability of electronic patient data offers health care practitioners increased opportunities for secondary use of EHR data to improve the quality of care and research [6-8]. However, the present literature does not describe the barriers that make the use of data and deidentification processes difficult nor does it focus on users' practical needs for data linking [9]. A study observed that "One of the fundamental steps in utilizing this EHRs data is identifying patients with certain characteristics of interest (either exposures or outcomes) via a process known as electronic phenotyping" [10]. Phenotyping is the process of extracting phenotypes from clinical data using computer-executable algorithms [11], and phenotypes are "the measurable biological, behavioural and clinical markers of a condition or disease" [12]. Phenotypes might be as simple as patients with type 2 diabetes or as complex as patients with stage II prostate cancer with urinary urgency but no indications of urinary tract infection [10].

There has been an annual rise at a rate of approximately 20% in primary care research using EHRs in the United Kingdom, which gathers data on general practice from the following databases [13]: Clinical Practice Research Data Link [14], The Health Improvement Network [15], QResearch [16], and Secured Anonymized Information Linkage (SAIL) [17]. However, with different data sets (eg, hospital, general practice, or emergency care), defining a condition is still very subjective, as there are many phenotyping algorithms for identifying the same condition (eg, there are currently 66 ways of defining asthma using routine health data) [18], and interpretation or manipulation of data often requires knowledge of complex programing languages, such as SQL [4]. This means that EHRs are still not accessible to many as their use requires specialized programing skills.

One of the most important factors for reproducible research is the availability of clinical codes in EHR-based research because researchers, clinicians, and health informatics professionals often use them to identify the target population and their specific conditions, known as phenotyping [8,19]. If researchers do not publish the code lists they used (eg, how they were established and the accurate phenotype definitions along with the original research using them), then an essential component of these studies is missing. In the absence of clinical code lists, data analysts would be unable to identify patients with or without conditions [19], and researchers would not be able to compare studies effectively. Even though code lists are available in some studies, researchers often encounter difficulties in retrieving relevant data from code lists created for another research project. Moreover, in specific uncommon conditions, minor errors in the selection of code lists may lead to misclassification of large numbers of patients, leading to biased results [20]. Although using previously developed phenotyping algorithms is often of interest to researchers in many studies, there are many challenges associated with reusing and replicating them effectively [21]. Therefore, it is extremely difficult to assess the validity and transparency of EHR-driven studies [22].

Although researchers request better transparency in sharing clinical code lists [23,24], they face difficulties in obtaining comprehensive code lists from EHR-based research. Although, there are currently no obligations from journals and funding parties to publish code lists, the Strengthening the Reporting of Observational Studies in Epidemiology and Reporting of Studies Conducted Using Observational Routinely Collected Health Data initiatives encourage transparency and open access to publicly available EHR-based research [25-27]. To address these challenges, different data linkage centers in the United Kingdom and other countries, such as Canada, have developed data portals for phenotypes (concept libraries), such as ClinicalCodes.org [22], Clinical Disease Research Using Linked Bespoke Studies and Electronic Health Records (CALIBER) data portal [4], and the Concept Dictionary at the Manitoba Centre for Health Policy [28]. Building web-based concept libraries enables data analysts, researchers, and clinicians to upload and download lists of clinical codes, update previous code lists, and share clinical code data across platforms, which would improve the validation of EHR-based research [22].

Objectives

This study aims to explore the needs of various users, including researchers, clinicians, machine learning experts, and managers, to develop a data portal for phenotypes (a concept library) and to examine why existing concept libraries are not widely used.

Methods

Design

A qualitative study using one-to-one interviews and a focus group discussion was conducted. We recruited a small purposive sample for in-depth one-to-one interviews in the first phase because it allows us to obtain substantial information from a small number of participants while also providing insight into their different viewpoints, needs, and experiences with concept libraries. In the second phase, we recruited a larger sample of participants for the focus group discussion to improve the generalizability of the results. The inclusion criteria were to recruit potential users of concept libraries from various disciplines, including researchers, clinicians, machine learning experts, and managers who conducted studies using routine data generated by data linkage repositories.

For this study, we adopted a semistructured approach. We created semistructured interview questions based on the Krueger and Casey format [29], which included introductory, flow, key, and final questions to be used in one-to-one interviews (Table 1). We also created a list of 10 questions based on the objectives of this study for the focus group session. The purpose of the questions was to generate thoughtful and thorough responses from the participants; therefore, closed-ended questions (eg,

yes or no) were avoided. The interviews and the focus group discussion were audio recorded and transcribed verbatim, and 2 thematic analyses were performed using the 6 steps of Braun and Clarke to identify the themes and subthemes [30].

Table 1. One-to-one interviews' questions guide.

Introductory questions	Follow questions	Key questions	Final questions
To improve repeatable research in Swansea, a team of developers is developing a prototype concept library. This is a portal that allows access to the read codes or International Classification of Diseases–10 codes to identify conditions. Do you think this will be a helpful resource? Is the concept library a good idea that we should continue to develop?	Do you know about other already existing concept libraries? What do you think about them? Something like this exists at UCL[a] called CALIBER[b]. Have you seen CALIBER? Have you used it?	• Do you prefer to use ready-made algorithms or to have access to them to modify them? • In your opinion, how should codes and algorithms be validated, and should they be validated? (Why should or should not?) • There are often different versions of a diagnosis (eg, highly specific and suspected or likely cases). Do you think we need to collect and validate the best two versions of a diagnosis (specific or suspected)? Or do you think we should put all possible methods of identifying a condition, valid or not, and allow the researcher to choose?	• What are your requirements for the concept library for it to be helpful and user-friendly? • What developments would you like to see to improve repeatable research using routine data?

[a]UCL: University College London.

[b]CALIBER: Clinical Disease Research Using Linked Bespoke Studies and Electronic Health Records.

Data Collection

The first author asked 6 participants from a variety of disciplines, including researchers (3/6, 50%), a clinician (1/6, 17%), a machine learning expert (1/6, 17%), and a senior research manager (1/6, 17%), at Swansea University and Cardiff University to participate in one-to-one interviews by email. The invitation email specified the aim and purpose of this study, the duration of each interview (30 minutes), and the location of the interviews, which might be their offices or a convenient and private location on the Swansea University campus.

Semistructured interview questions, which follow the structure proposed by Krueger and Casey [29], were used (Table 1). The structure of the interview questions consisted of introductory, flow, key, and final questions. The purpose of the introductory questions was to help the participants talk freely about their overall experiences. The flow questions were designed to create a smooth transition to the key areas that the authors intended to explore. The final questions were designed to summarize the interview and ensure that the participants did not have further comments [31].

Before conducting the interviews, the first author explained the purpose of the research and what it involved, and at the beginning of each interview, participants received additional verbal and written information about the research project. The interviews were conducted at Swansea University Medical School in a place selected by the participants (eg, their office). After 5 interviews, no new themes were observed and interview 6 confirmed that no new themes emerged. The interviews were audio recorded and transcribed verbatim. Thematic analyses were then performed using the 6 steps of Braun and Clarke to identify the themes and subthemes [30].

All researchers working with the SAIL databank, a national eHealth data linkage infrastructure in Wale (N=34) were invited by email to participate in the focus group discussion, and 14 (14/34, 41%) researchers attended the focus group discussion. In total, 2 focus group discussions, each of which had 7 (7/14, 50%) participants, were held for 2 hours by 2 moderators (ZA and SB), who used the same set of semistructured questions to perform a SWOT (strengths, weaknesses, opportunities, and threats) analysis for the current system for phenotyping and the proposed concept library. We used a SWOT analysis tool in this study because it enabled the participants to discuss what they liked (strengths), what advantages would be gained (opportunities), and what problems (weaknesses) and issues (threats) they felt needed to be tackled. Although the 2 moderators used the same set of questions, the order of the questions was adjusted to the needs of each group.

At the beginning of the focus group discussion, the first author gave a brief presentation about concept libraries, including defining concept libraries, explaining their potential uses, and mentioning examples of some of the existing concept libraries in the United Kingdom. A second presentation about the Swansea University prototype concept library was then given by one of its developers. Feedback from the participants was sought concerning their perceptions of the concept library's needs and their evaluation of the strengths and limitations of the proposed concept library. Participants' perceptions of existing concept libraries, as well as their assessment of the proposed concept library's strengths and limitations, were explored using the following set of semistructured questions:

- What are your thoughts regarding the proposed data portal for phenotypes (a concept library) when it rolls out?
- Do you think this is worth doing? Would you value this?
- Has anybody used existing concept libraries? What have you experienced with them?

Let us talk now about your current system for phenotyping:

- What do you do? What are your methods?
- Are you happy with them? Or what would you like differently?
- What are your thoughts on this plan (building a concept library)?
- Would you use it? Would you share your phenotypes and your phenotyping algorithms?

If you do not want to share your work:

- Can you tell us why? And what motivates you to share it with others?
- Of all the things we have discussed, what is most important to you?
- Is there anything we should have talked about but did not?

The goal of using the SWOT analysis was to identify positive factors that operate together and the potential difficulties that must be identified and solved. During the focus group discussions, participants expressed their own opinions and listened to the opinions of others. As the discussions progressed, participants began to ask questions of one another and share similar experiences. This increased the depth of the conversation. The SWOT analysis gave us a full picture of views and experiences of concept libraries by the participants, making this a holistic evaluation with the ability for participants to hear and comment on each other's responses. Textbox 1 presents a summary of the SWOT analysis in the current system for phenotyping and the proposed concept library. The 2 focus group discussions were audio recorded and transcribed verbatim. Thematic analyses were then conducted using the 6 steps of Braun and Clarke to discover the main themes and subthemes (Table 2).

Textbox 1. A summary of a SWOT (strengths, weaknesses, opportunities, and threats) analysis of the current system for phenotyping and the prototype concept library.

SWOT analysis

Strengths

- Concept libraries provide researchers with a good starting point.
- Publicly available code lists may provide researchers with a history of a particular area of research, such as asthma.
- Referencing previously published lists of codes enables researchers to demonstrate a rationale for using such lists of codes.
- Using research methods developed by others that match the researchers' interests could result in significant time saving.
- Collaboration among researchers is facilitated through sharing and using research methods such as code lists.

Weaknesses

- Searching for and reusing phenotypes and codes is a time-consuming and labor-intensive process.
- There are various lists of codes for each phenotype definition.
- The list of codes chosen by clinicians varies significantly.
- A large number of previously developed code lists could not be repeated.
- Reusing other researchers' data requires programing knowledge such as SQL.
- Some of the ready-made phenotyping algorithms may not be very useful in terms of their general purpose.
- Some existing concept libraries have limited user interfaces.
- Some existing concept libraries are not user-friendly.
- It is unclear who is accountable for the quality of the uploaded codes in concept libraries.
- The validity of the content of concept libraries is unclear.

Opportunities

- Concept libraries must provide user documentation.
- Concept libraries must provide users with training.
- Transparency in sharing the whole approach used to create the code lists is required.
- Establishing a standardized way of defining each specific condition to facilitate comparisons of research outcomes across the United Kingdom.
- Creating a specialized library that stores code lists of a specific condition within a specific set of patients, such as a concept library specializing in chronic conditions in children.
- Creating a concept library that engages a wide variety of users (ie, is easily understandable by clinicians but has some advanced features such as programing skills for more expert users).

Threats

- The inconsistency of data across various databases makes data reuse difficult.
- Lack of confidence in the quality of the list of codes developed by other researchers if they are not cited.
- Access to code lists is limited as some researchers do not publish them alongside their studies.
- Different research outcomes result from a lack of access to a list of codes created by other researchers.
- Data sharing may be inhibited if there are no returns, such as referencing and acknowledgment.
- Concerns about ownership rights discourage data sharing (eg, methods could be used as their own by other researchers before publication).

Table 2. Presentation of the themes and subthemes of the one-to-one interviews.

Themes	Examples of participant narratives
Theme (1): previous opinion of a prototype concept library	
Positive	"If there's a way of doing that already that is set up and is validated and is consistently applied that would be an amazingly useful resource" (researcher 2).
Neutral	"It will be helpful, but it needs to be extended. If they want to build something like this, and it is effectively working as a library, you need two things to be happened: (1) people are happy to feed in their constructs so it builds up, and (2) a useful library, easy to go, to browse, and to borrow phenotypes definitions" (a clinician).
Negative	None
Theme (2): requirements of a prototype concept library	
Usability	
Simplicity	"Simple plain English not in SQL or python" (a clinician).
Searching ability	"What is the type of search engine? Is it a search engine that just does disease phenotypes or also does the health status phenotypes or risk factor phenotypes, symptoms phenotypes?" (a clinician).
Data quality	"It's really just about transparency and documentation. So, anybody can effectively do anything that can be turned into a reproducible research output. The barriers are usually not enough time to comment and document it properly and then not enough quality assurance" (a senior research manager).
Sharing ability	"It would be very useful to share the knowledge about codes such as read codes, ICD 10 codes, or OPCS codes, and share ideas and concepts between other users that will save lots of time" (researcher 3).
Sustainability	
Interoperability	"How interoperable it is with other systems because the major failure of most of these systems is that they're not interoperable, so people don't use them" (a senior research manager).
Accessibility	"So, from a group like myself, or me as a user, we would probably like direct access to the underlying data it stores. So, whether that's through something like SQL directly, or something like that through a statistical package, because where we do lots of bulk type work" (a senior research manager).
Analyzability	"I wanted to look at all health codes of my study population. Then, through machine learning, like feature selection, I tried to identify the most important list of codes, which are associated with the popular health conditions" (researcher 1).
Theme (3): user experience of existing concept libraries	
Aware (used them)	"Yes, so with QOF, we definitely used QOF codes a lot, because obviously going back to the quality assurance question, they'd been assured so that the NHS can use them for remuneration of money and payments. With other systems, we tend to look online to see CALIBER of things with us, then yes we have used outputs from those systems before" (a senior research manager).
Aware (not used them)	"No. I have not used any of these things before so I think there is CALIBER and I think, is that part of what was set up within the previous Farr institute? so I am aware that some of these exist but I haven't looked into them before" (researcher 2).
Not aware	None
Theme (4): user's recommendation to improve repeatable research	"If we want reproducible research, we have to all be using these resources in a similar way or at least we need to be able to understand what previous projects have done. It is about setting things out clearly. Clear definitions, clear sets of codes that people can then either use themselves or build on I think" (researcher 2).

Data Analysis

The interviews and the focus group discussion were analyzed separately following the analysis approach by Braun and Clarke [30]. The transcripts of the interviews and the focus group discussion were read several times, and then the initial codes were grouped into themes and subthemes using a qualitative data analysis software (NVivo, QSR International) [30,32]. ZA had read all the transcripts, and SB read a sample of the transcripts. They independently identified the themes and subthemes, then met regularly to compare them and reach an agreement on what was being done. Themes and subthemes were discussed with respect to their relevance to the research question in the data collected. They critically reviewed the themes again to determine their primary meanings, and similar initial themes were combined into one theme. They discussed the definitions of the relevant themes in the research questions and applied appropriate names to describe each in this study. Textbox 2 provides further description of the thematic analytic steps.

Textbox 2. The 6 thematic analytic steps used for this research.

Thematic analytic steps

Self-familiarizing with the data

- ZA transcribed half of the audio recordings from the interviews (3/6, 50%). The other half of the audio recordings from the interviews (3/6, 50%) and the audio recordings from the focus group discussion were transcribed by professional transcribers. During this phase, ZA read all the interview and focus group discussion transcripts several times, and SB read samples of them. ZA and SB considered all the topics discussed by the participants, recorded notes on these topics in the transcripts, and then organized them in a note book.

Creating initial codes

- After familiarizing themselves with the data, ZA and SB worked independently to identify initial codes from the transcripts that summarized what was said during the interviews and focus group discussion. They organized the identified codes into meaningful groups using qualitative data analysis software (NVivo, QSR International). They used the same coding procedure for all the transcripts.

Searching for themes

- ZA and SB started interpreting the initial codes using their extracted data, and they began grouping the codes with similar meanings together. Using the NVivo software (QSR International), the initial codes were then sorted and labeled into themes and subthemes depending on the meaning or relations shared by the codes.

Revising themes

- ZA and SB critically reviewed and refined themes against the data several times to determine their core meanings, and similar initial themes were combined into one theme. To reach an agreement, themes and subthemes were discussed in terms of their relevance to the research question.

Defining themes

- Each of the themes identified in the previous steps was named and defined by ZA and SB. They used the initial labels created for the themes to provide appropriate names that describe the meaning of the themes in this study. ZA and SB defined each theme based on the content and meaning of their codes, and they examined these definitions in relation to their relevance to the research questions.

Writing up the report

- After defining and naming the themes, ZA and SB began writing the findings for this manuscript. They used quotes from the participants' responses that related to the themes and the research question to illustrate the findings.

Ethics Statement

Ethical approval to conduct the research was approved by the Research Ethics Sub-Committee of Swansea University, project reference number 2019-0007.

Results

Interviews With Users

Overview

In total, 6 one-to-one interviews were conducted, and each interview lasted for approximately half an hour. The analysis of the interviews resulted in 4 main themes, with several subthemes (Table 2). The four main themes are as follows:

1. Previous opinion of a prototype concept library
2. Requirements of a prototype concept library
3. Experience of existing concept libraries
4. Recommendations to improve repeatable research

Previous Opinion of a Prototype Concept Library

The majority of the participants were positive about the prototype concept library and felt that a concept library in principle was a very helpful resource for conducting repeatable research. A machine learning expert mentioned that a concept library will be an extremely useful resource because read codes from general practice and International Classification of Diseases (ICD)–10 codes from hospitals are the most common data items that machine learning experts would like to use most often. They use data linkage repositories to extract the necessary data for machine learning in public health studies, and they use the codes to extract the data from the repositories. Researcher 3 said, "It would be very useful to share the knowledge about codes such as read codes, ICD 10 codes, or OPCS codes, and share ideas and concepts between other users that will save lots of time. It is useful to use verified codes," and researcher 2 stated, "If there's a way of doing that. Already that is set up, and is validated and is consistently applied, that would be an amazingly useful resource."

However, 2 participants (a clinician and a senior research manager in health data science) were not sure about the effectiveness of the prototype concept library because they felt that users had to engage with it for it to be useful and they were not sure how well users would engage: "There is potential that it could be useful as a tool. It will kind of come down to how usable it is, how flexible it is, how well it's maintained, how much of the community uses it" (a senior research manager).

Requirements of the Prototype Concept Library

The participants mentioned several requirements they would like to see in the prototype concept library. For example, they stated that the concept library needed to have high usability. This means that it needs to be simple and easy to use by naïve users: "It should be simple enough, within one or two clicks;

we can find the required data, but also should contain advanced expert features (R, SQL, or Python programing languages) to extract, include, or exclude codes necessary for their studies" (researcher 3) and "Like, in one of my previous projects, I looked at, from a machine learning perspective, I wanted to look at all health codes of my study population. Then, through machine learning, like feature selection, I tried to identify the most important list of codes, which are associated with the popular health conditions" (researcher 1). They also stated that the concept library should have a good search engine so that they can easily find the phenotypes and phenotyping algorithms they want to use. A clinician inquired, "What is the type of search engine you are developing? Is it a search engine that just does disease phenotypes? or also health status phenotypes, risk factor phenotypes, or symptom phenotypes. For example, I am looking for diabetes, but I may also be looking for smoking or alcohol consumption, or symptoms like pain or cough. So, how big is the enterprise and how do you search for what are the appropriate terms? Discussion is needed to know what is it?"

In addition, the participants stated the following requirements:

1. Include the data sources used (eg, codes from general practice, hospital [ICD and Systematized Nomenclature of Medicine], and British National Formulary medication), a general clinical code list for comparison, lists of ontologies along with their variances and versions, and a description of how codes were established: "It is about setting things out clearly. Clear definitions, clear sets of codes that people can then either use themselves or build on, I think" (Researcher 2).
2. Have a clear phenotyping algorithm labeling convention for search engines. A clinician stated,"What do you search on? Thought about what do you call these phenotypes? Is there a consistent in calling them? For example, Type II diabetes, or insulin dependent diabetes" and researcher 1 stated, "So, first of all, for the code reference library, two things are always there in my mind. It's in my opinion again. Number one, they should be validated. Secondly, they should be correctly labelled."
3. Specify why a particular phenotyping algorithm was developed (eg, definite disease or probable/suspected condition definitions): "When I have an algorithm, I want a field that tells me the purpose of the algorithm, a brief description of what the algorithm is intended to do" (a clinician).
4. Illustrate the logic model category used to create phenotyping algorithms (ie, code lists, inclusion or exclusion factors, and clinical or machine learning approach used). "Is this just a code list of inclusion factors? And or exclusion factors? Or is it static? Does it have a tampered relationship? So, some algorithms are present or absence of conditions, some required a tampered dependence. In the logic model categories: Is this a clinically derived algorithm from experts' views or for instance that machine learning derived algorithms" (a clinician).
5. Use ready-made phenotyping algorithms that can be modified to fit the needs of their research. All participants agreed that if they had to create their own phenotyping algorithms because ready-made phenotyping algorithms could not be modified, they needed an easy approach to use a code list in the concept library.

There was an issue regarding how to validate phenotyping algorithms, and most participants expressed their preferences for using all possible methods of identifying a condition, valid or not, to allow the researcher to choose the phenotyping algorithms according to their research requirements: "So, there is no right answer for that because it's going to be very dependent on your research question, your study group, and your study design. So, once again, if the concept tool is going to match multiple different use cases, it's going to need to accommodate for those different types of study design" (a senior research manager). Sharing phenotyping algorithms needed to be easy and not time-consuming, and some felt there needed to be some recognition of their work before they would give their codes. Finally, a concept library must be interoperable with other products or systems: "How interoperable it is with other systems, because the major failure of most of these systems is that they're not interoperable, so people don't use them" (a senior research manager). Most participants wanted the source code (eg, the SQL code for the phenotyping algorithm itself) to be available in a downloadable machine-readable format to be able to access it using specific programing languages such as R, SQL, or Python.

Experience of Existing Concept Libraries

All participants stated that they were aware of some existing concept libraries, such as CALIBER and ClinicalCodes.Org (both in the United Kingdom), but most of them did not use them. The reasons given for not using them were that they already had their own self-made concept libraries (eg, concepts they have used before) or the available concept libraries did not provide phenotyping algorithms that fit their studies. For example, a machine learning expert mentioned the reasons for not using two of the existing concept libraries, namely the Concept Dictionary at the Manitoba Centre for Health Policy in Canada and CALIBER in the United Kingdom were, "Canadian systems provide Canadian data for their studies, CALIBER is specific for cardiovascular disease and does not have many concepts in it." Conversely, 2 of the participants mentioned that they used some existing concept libraries to extract and develop phenotyping algorithms for their studies: "We definitely used QOF codes a lot, with other systems, we tend to look online to see CALIBER, we have used outputs from those systems before" (a senior research manager).

Recommendations to Improve Repeatable Research

The participants suggested the following recommendations to improve repeatable research output using routine data:

1. There should be a drive for more transparency in research methods documentation, such as publishing complete phenotype definitions and clear code lists. A senior research manager stated, "It's really just about transparency and documentation. So, anybody can effectively do anything that can be turned into a reproducible research output," and researcher 2 said, "If we want reproducible research, we have to all be using these resources in a similar way or at least we need to be able to understand what previous

projects have done. It is about setting things out clearly. Clear definitions, clear sets of codes that people can then either use themselves or build on, I think."

2. Providing opportunities for researchers to collaborate rather than working in isolation, "The barriers are usually not enough time to comment and document it properly and then not enough quality assurance. So, if there was more time and or more availability of those kinds of opportunities for people to collaborate rather than doing things in isolation, there's almost all the research we do here could be turned into a reproducible type of output" (a senior research manager).
3. Develop a concept library that enables researchers to begin classifying population outcomes using uniform codes: "I think that a resource like this is a very good step in the right direction because I think what people need to start doing is using consistent codes in order to identify conditions or outcomes within populations" (researcher 2).
4. Provide validated phenotyping algorithms that researchers can use directly to avoid duplication, with the ability to modify them to meet their own research needs: "For each project, it always has some specific requirement which is unique, which is not common. There are some things which are common, and there are a few things which are very unique. So, we need to have some algorithms which we can just use to, you know, just to avoid the duplication, but also, we need to have control of the algorithms, so that we know only that these bits are going to be different for this project, so I'm going to replace, change, modify this bit, and we'll run it" (researcher 1).

Focus Group Discussions

Overview

Of the 34 invited researchers, 14 (41%) attended the focus group discussion. These participants were researchers (14/34, 41%) from Swansea University who were working with the SAIL data in the Data Science Building. Of the 14 participants, 5 (36%) were female participants and 9 (64%) were male participants. Furthermore, 6 (43%) participants were PhD holders, 6 (43%) were Master's degree holders, and 2 (14%) were Bachelor's degree holders (Table 3).

Table 3. A summary of general information on the participants in the focus group discussions (N=14).

Parameters	Information
Current job position, n (%)	• Data scientist, 13 (93) • Financial planner, 1 (7)
Sex, n (%)	• Female, 5 (36) • Male, 9 (64)
Education, n (%)	• PhD degree, 6 (43) • Master's degree, 6 (43) • Bachelor's degree, 2 (14)

Research interests	• Data scientists
	  ◦ Concept libraries
	  ◦ Repeatable research with large health data
	  ◦ Phenotyping and code lists of cancer disease
	  ◦ Respiratory disease
	  ◦ Algorithm or reusable codes development
	  ◦ Asthma
	  ◦ Collaboration in research methods
	  ◦ Data analysis
	  ◦ Machine learning
	  ◦ Arthritis
	  ◦ Health informatics
	  ◦ Musculoskeletal disorders
	  ◦ Healthy aging
	  ◦ Gut—brain axis
	  ◦ Neurodegenerative conditions
	  ◦ Statistical methods
	  ◦ Epidemiology
	  ◦ Cancer
	• Financial planners
	  ◦ Intervention between primary care and secondary care and how they interact

The focus group discussion was held for 2 hours to perform a SWOT analysis of the current system for phenotyping and the proposed concept library, which was recorded and transcribed, and thematic analysis was conducted on the transcripts, which resulted in the identification of the following seven main themes:

1. Facilitators for and barriers to participants' contributing their research methods
2. Facilitators for and barriers to participants' use of other researchers' methods
3. Participants' concerns about the prototype concept library

4. The requirements of the participants for the prototype concept library
5. Participants' recommendations to improve repeatable research
6. Participants' perceptions of their current phenotyping system
7. Participants' use and perceptions of existing concept librariess

Facilitators and Barriers to Participants' Contributing Their Research Methods

Facilitators

Several facilitators were identified by participants as motivators for them to share their work (eg, phenotyping algorithms and code lists). Many participants stated that being credited appropriately (eg, receiving citations from other researchers) would motivate them to share their work: "If whoever's using it acknowledges it's use in whatever they publish, at least you're getting some recognition " (data scientist 8) and "If there were DOIs attached to the code list of algorithms, when people are publishing, there's an incentive for putting it on there, because they're able to demonstrate the impact their work has had" (data scientist 4).

Some participants stated that communicating with their research team would encourage them to organize team resources and discuss research findings from other researchers who used their code lists. However, improving research opportunities, increasing academic achievement, and sharing knowledge through collaboration with other researchers working in the same organization would motivate some of the participants to share their work: "I think there's benefit to the organization, and there has to be benefit to the people contributing to it" (data scientist 4). In general, researchers work in an organization (eg, a university or a research institute), and they work hard to improve the research outcomes of their organization. Some participants stated that advancing the research base and saving other researchers' time and effort would stimulate them to share their work: "Surely if you've done something you think really worthwhile, you want other people to use it, as well, because then that furthers the research" (data scientist 6).

Barriers

On the other hand, the participants pointed out several barriers that could inhibit them from sharing their work (eg, phenotyping algorithms and code lists) with other researchers. Some participants argued that it is easy to build a phenotyping algorithm that fits exactly their needs, but it is more challenging to develop a general one, so it can be used by others (eg, many clinical researchers have created phenotyping algorithms for particular research, and these algorithms are difficult to generalize).

Several participants mentioned that a lack of return for their hard work (eg, not receiving any credit from others, such as referencing when they reuse their data) would prevent them from sharing their work: "How do you enforce that people are going to give you credit? It doesn't happen sometimes, when referencing, saying where they got it from. You've just got to hope they do" (data scientist 11). Some participants were worried about their intellectual rights (eg, if they shared their methods such as phenotyping algorithms before publication, other researchers would use them as their own).

Facilitators and Barriers to Participants' Use of Other Researchers' Research Methods

Facilitators

The participants mentioned several facilitators that would encourage them to reuse research methods developed by others, such as the following:

1. Using existing code lists can save them a lot of time and effort, which they frequently spend creating new code lists from scratch: "It's the first stage of every single process, and we tend to get two or three months of work, until we get to that final code list, and we can now start looking at the cases" (data scientist 10).
2. Reusing available data, such as code lists, is a good place to start for researchers (for example, they can use them to examine new ideas and gain new insights):"Having code lists would be such a help, to get you started. They always want things like BMI and weight and height. There are hundreds of codes for those. The smoking codes, having a list, even if you don't use the algorithm that they've developed, is a huge bonus" (data scientist 12).
3. Using the work of others as a reference to compare research outcomes, and researchers want to prove that there is a basis for the use of such codes.

Barriers

Conversely, the participants pointed out several barriers that could inhibit them from reusing methods developed by other researchers such as the following:

1. Poor data quality discourages researchers from reusing it: "You could upload complete garbage" (data scientist 1).
2. Some phenotyping algorithms will not work outside the population in which they were developed. For example, code developed in Canada may not be relevant to finding conditions in general practitioner data in the United Kingdom: "Yes, it works in their population, because where they've trained it." (data scientist 5). .
3. Whether the data are useful to researchers plays an important role in the decision to reuse them (eg, researchers would not use a phenotyping algorithm if its general purpose did not match their interests): "Yes, a general-purpose algorithm may or may not be very useful to have it to see what they've done, but you may not use it" (data scientist 12).

Participants' Concerns About the Prototype Concept Library

When researchers decide where to deposit, share, and reuse data, they prefer to use approved concept libraries: "Is it going to be approved?" (a financial planner). Moreover, some participants stated that it is not clear who is responsible for the quality of the phenotyping algorithm, if this is the responsibility of the developers running the concept library or the responsibility of the researchers uploading the phenotyping algorithms: "If people send the codes, the onus of the quality

of that code list you would still want to be the responsibility of the researcher to be submitting worthwhile codes. You don't want to then be the guardian of the quality of the code list. You still need to know where the responsibilities lie" (data scientist 4). Researchers do not want to upload phenotyping algorithms if they could be *blamed* for flaws, and health informatic developers do not want to take responsibility for the phenotyping algorithms that were uploaded.

The participants expressed their concerns about the completeness rate of the phenotyping algorithms. They would like to know the percentage of the gap to be considered when using a phenotyping algorithm from the prototype concept library: "What is the completeness rate? For certain things, we know there are gaps. If the gap is 20%, is that something I should be including in any algorithm I'm considering?" (data scientist 8). In addition, there has been a question as to whether codes need to be peer reviewed so that quality is evaluated.

Requirements of the Participants for the Prototype Concept Library

Usability

1. Learnability: Some participants said they would like the concept library to be easily understandable by clinicians, who acknowledge the clinical definition of the code lists with little technical skills to simply point and click the selected code lists, whereas other participants requested the availability of advanced functions to be used by expert users: "The concept library should be easy. Someone needs to train us" (data scientist 9).
2. User documentation: A collection of well-defined task-oriented documentation for users was required by some participants. They want a user documentation that consists of clear, step-by-step instructions on how to use the concept library and gives examples of what the user can see at each step (eg, screenshots would be useful): "Concept library should have some documentation" (data scientist 9).
3. Data quality: Some participants required the availability of a consistent method for identifying each specific condition to ensure that what researchers are doing is compatible within their immediate team but also within the broader research community in the United Kingdom to facilitate a comparison of research outcomes. Other participants stated that they needed a predefined list and a uniform approach describing how to use existing codes of additional diagnoses, such as smoking: "Additional things like smoking and alcohol status are used a lot, but they're usually very different for every project. We should have a more uniform way of doing it, like, we'll take that bit off the shelf and use it, and do the bespoke bit for things that need to be bespoke" (data scientist 5). If there are multiple code lists for the same condition, some participants proposed that versions be generated to describe each particular condition: "So, it would be relevant that there were multiple lists for the same condition, if you've got a version and way of defining a certain condition" (data scientist 4).
4. Transparency: Several participants required transparency in sharing the entire approach used in developing the code lists, including phenotyping algorithms and the methods used. They stated that if they use a code list for each comorbidity of a condition, they will build an entirely different score over the years. Therefore, transparency in the documentation of research methods would help them to know which score is the best.

Sustainability

1. Accessibility: Several participants needed the availability of an access control that allows access to the codes only after publication, while at the first stage of the study, researchers spent a lot of time and effort developing them, and they feared someone else could publish work faster than them using the algorithms: "There should be an option in the concept library for lists that have been published. People can develop them, but if they're not published, you don't have to use them" (data scientist 3).
2. Licensing: Some participants needed to know which type of license was adopted by the developers of the concept library (eg, researchers can have one that means any researcher can take it and use it, or they can have one that means researchers can use it but not for commercial purposes).
3. User community: Several participants required users to quote a reference if publishing papers based on the results (partially or completely) derived from the concept library: "If I want to use someone else's work, I think that's the norm, and should be in this economy. Anything, not just code. To use this, I should reference that it's based on this or other thing completely, or a part of it" (data scientist 2). Referencing helps to determine whether there is or will be an active user community for the concept library and the codes used: "It potentially would make your publication more discoverable. If there's a whole community of users using this" (data scientist 1).

Participants' Recommendations to Improve Repeatable Research

Of the 14 participants, 9 (64%) suggested that the prototype concept library should be accessible both in the United Kingdom and globally and practically available to enable researchers around the world to use an web-based secure platform, which stores codes and other logic, and to encourage researchers to contribute their codes to promote research: "Should be open for the United Kingdom" (data scientist 9). However, a participant recommended that the prototype concept library should be closed at the beginning to ensure it is working and then to become opened as researchers build trust: "You might need to restrict it, to start with, to make sure it works. Otherwise, everyone will see the problems you might have" (data scientist 12). In addition to know who is using the concept library, data scientist 8 suggested that it should have request sharing followed by open sharing.

Accessibility to research data has significant potential for scientific advancement as it promotes the replication of research results and enables the use of old data in new contexts. With respect to this, some participants suggested that funders and publishers should obligate researchers to share their research data such as code lists: "Some sort of obligation by funders to

share this" (data scientist 2) and "Publishers, as well" (data scientist 8).

A participant suggested the use of preauthorization of publication by journals based on the research protocol because researchers can put their protocol first, and all the limitations are actually corrected before they run the research. This approach has many advantages for both the researcher and the publisher, as it improves the quality of the output. Another participant recommended the creation of a discussion forum in the concept library to facilitate collaboration among researchers on just about any topic (eg, they can share their ideas, submit their comments, and discover new ideas): "Make it almost a forum" (data scientist 8).

Participants' Perceptions of Their Current Phenotyping System

The participants mentioned several problems associated with the current phenotyping system. For example, they have to search for codes from different databases, which use different coding systems such as read codes and ICD-10 codes, and then they have to validate the selected code lists with experts in the field such as clinicians: "I have to google all of this and search what was there within the community. I have to go to CALIBER, I have to go to Manchester, or there is a work in Edinburgh University, do some work there. Do the search. I have to go there, see the ability to work, and start. It does take a lot of time. Based on my study of Google, I have to start a record, and I have to validate it, verify with other people, clinicians or researchers. It's a long process" (data scientist 9).

Although they could find some codes on the web, they still had to locate the list manually, copy it, and enter the codes into their scripts. Often, they might spend a few days on it, and they might miss obscure codes or even use irrelevant codes: "Starting from scratch, I would go online to see what's available. Go into other people's and see their code lists" (data scientist 11). With respect to this, some participants said that they preferred to use code lists that were referenced or used by other researchers.

Some participants reported that the read code lists chosen by the researchers were different from the read code lists chosen by general practitioners. For example, they found that there were some very clear codes, but they were rarely used by general practitioners: "What we get in the read code list isn't necessarily what the GPs are recording it under" (data scientist 12). They also stated that there is a significant difference between what one general practitioner may say in a list of codes versus another: "For example, there is no single entity code for asthma. There are different entities. If you want to find specific things within asthma, there's a list of codes for them" (data scientist 2).

Participants' Use and Perceptions of Existing Concept Libraries

Not all participants had previously used some of the existing concept libraries. However, most of those who used some of them expressed negative perceptions. For example, several participants stated that the concept libraries they used were not user-friendly (ie, they were difficult to use by new users): "For CALIBER, it seems not so user friendly. It's not easy. You have to know first. Someone needs to train you up. For new users, it's difficult to get inside CALIBER. The concept library should be easy. Someone needs to train us. Concept library should have some documentation" (data scientist 9). Therefore, training and good user documentation are required. A further problem for some participants was the inconsistency of data among various databases, which makes reuse of data quite challenging. "But if there is something that gets secondary and primary care involved, and there's a registry, if the definitions that are created in Manchester, how easy will it be to apply it to, for example, in Wales or Scotland, where registry is a bit different?" (data scientist 8).

Participants who did not use any of the existing concept libraries expressed different perceptions about them. For example, some participants reported that they wanted to explore available concept libraries. Others, however, expressed doubts about the quality and validity of the data stored in these concept libraries, which could prevent them from using them: "I haven't looked at them myself, but if you go on this clinical code site and you type in diabetes, there are 50 different code lists people have put together for diabetes" (data scientist 6). Some participants stated that the main reason for not using any of the existing concept libraries is not finding a concept library that matches their studies. The developers of concept libraries may consider building a specialized library that stores code lists of a particular condition within a specific group of patients according to researchers' needs, such as developing a concept library that specializes in chronic conditions in children.

Discussion

Principal Findings

Development of a concept library that meets users' expectations is extremely useful for repeatable research (eg, researchers would be able to use archived code lists to compare studies). This study found that, although in principle, everyone felt that a digital portal containing a concept library would be very helpful, there were many requirements needed before its development. It needs to engage a wide variety of users if it is to be used (and current concept libraries are not widely used), which means that it has to be very simple (point and click) for some, but it should have the software and usability to manipulate and design phenotyping algorithms for more advanced users. In addition, it needs to have a very high-quality search engine so that it is very easy to find information, and for it to expand, there needs to be a reason for users to upload their phenotyping algorithms, which need to be very easy and quick.

This study indicated that although most of the interviewees expressed positive impressions about the idea of building a prototype concept library, approximately half of the participants expressed an interest in contributing to it. For the prototype concept library to work, researchers must engage with it and upload their codes there so that other people can use them. If researchers did not share their codes in the prototype concept library, this would usually mean an empty library. For better adoption of the prototype concept library, it is recommended that the developers consider the various facilitators for and

barriers to participants sharing their work and reusing the work of others.

The findings of the focus group discussion demonstrate that facilitators for the participants' sharing of their research methods vary across four categories: (1) personal drivers (eg, obtaining appropriate credit, such as citations)—this confirms the results of earlier studies that suggest that researchers may be motivated to share their work if sharing leads to an increase in their citations [33-35], (2) benefits for their research team (eg, sharing information to promote research within their team) [36,37], (3) benefits for their organization (eg, collaboration among researchers working within the same organization would advance their organization's research outcomes), and (4) benefits for the research community (eg, expanding research base) [38]. With respect to this, Cragin et al [39] have stated, "As a research group gets larger and more formally connected to other research groups, it begins to function more like big science."

There were several barriers that could inhibit the participants from sharing their research methods, such as the expected performance of the shared methods (eg, they felt that building a general phenotyping algorithm to be used by others is very difficult) [40] and lack of personal benefits such as recognition (eg, they were worried about not being referenced by researchers who used their methods). In relation to this, Molloy et al [41]. reported that researchers can be discouraged from sharing their work by fear of not obtaining sufficient credit. Therefore, a safeguard against uncredited use is necessary [42]. In addition, participants mentioned that they were afraid that their methods would be used by other researchers as their own before publication. The results of the study conducted by Huang et al [43] indicated that although most participants are interested in sharing papers related to biodiversity data, >60% of the participants were reluctant to share primary data before publication. Moreover, findings from this study correspond with other studies regarding the need to adapt impact metrics to promote data sharing [44,45] because researchers would not be able to measure the success of their methods if metrics are not available. Unless these obstacles are resolved, the sharing of data in concept libraries is unlikely to increase significantly.

Several facilitators encouraged participants to reuse research methods developed by others. They reported that reusing code lists created by other researchers would make their task much easier, save them a lot of time, and help to demonstrate that there is a justification for using such codes. These findings are consistent with those of the previous studies. For example, Anneke and Helen reported that researchers are using open research data to "be aware of the state of the art and not recreate the wheel, as well as access to more data and generating fresh insights" [46].

The results of this study indicate that more than half of the participants were not satisfied with their current system for phenotyping for several reasons, including the lack of accessibility of other researchers' work, such as code lists, which could affect research outcomes and the fact that reusing publicly available code lists consumes a lot of time and requires lots of work [38]; lack of confidence in web-based code lists if they are not cited by other researchers; lack of availability of a consistent approach for defining covariates such as smoking; and the selected read code lists by the researchers are different from the selected read code lists by the general practitioners. It seems that their current approach lacks confidence and is time-consuming and effort-intensive.

This study demonstrates that existing concept libraries are not widely used, and most participants who used some of the existing concept libraries expressed negative impressions about them (eg, they do not provide training or user documentation, and they are difficult to use) [36-38]. Lack of knowledge of the existence of concept libraries and how to use them is generally described as an obstacle to data sharing [47]. As existing concept libraries are not used by all researchers, obstacles that inhibit researchers from using them need to be addressed when building new concept libraries.

Strengths and Limitations

To our knowledge, this is the first study aimed at identifying the needs of various users of a concept library. The findings of this study would have a significant impact on improving the efficiency of existing concept libraries by informing their developers about the different requirements, facilitators, barriers, and recommendations of the various users. In addition, this work will greatly inform the developers of new concept libraries to improve access to and collaboration with EHRs' routine data, which is part of an all-UK agenda, and the findings of this study will have implications for other countries working to access and share EHRs' routine data.

This study has some limitations that should be addressed in future studies. The first limitation is that we had a time limit on how long we could talk to the participants because each one-to-one interview was given 30 minutes. As a result, the number of questions we could ask and the amount of time we could spend on each question were limited. The second limitation is that all the participants of the interviews and focus group discussion were recruited because they used the SAIL databank, a national eHealth data linkage infrastructure in Wale, so they mostly talked about the Swansea concept library in the SAIL databank. As the discussion focused on the SAIL databank, its generalization to other concept libraries was limited.

Conclusions

In conclusion, although it may seem beneficial for researchers to reuse methods developed by others, such as code lists, some researchers who created them prefer not to share them because they worked hard to create them and would rather publish them first to ensure their academic rights, such as being referenced [48]. The major challenge is that some researchers would like to use the work of other researchers, but they do not want to contribute their work to concept libraries. Open sharing can be more difficult in the research community as researchers compete for grants, work promotions, and publication quotations [48]. They think carefully about how, when, and where to share their work as they have spent a vast amount of time and effort to develop it [47]. A solution to these issues would be to encourage researchers to contribute data to the prototype concept library in such a way that the shared data is understandable and reusable

(eg, ensuring uploading of adequate documentation) for the public good rather than for personal gains.

Acknowledgments

This research was supported by the Kuwait Cultural Office in London, HDRUK (Health Data Research UK), and the National Centre for Population Health and Wellbeing.

Conflicts of Interest

None declared.

References

1. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, MIT Critical Data Conference 2014 Organizing Committee. Making big data useful for health care: a summary of the inaugural mit critical data conference. JMIR Med Inform 2014 Aug 22;2(2):e22 [FREE Full text] [doi: 10.2196/medinform.3447] [Medline: 25600172]
2. Wei W, Teixeira P, Mo H, Cronin R, Warner J, Denny J. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016 Apr;23(e1):20-27 [FREE Full text] [doi: 10.1093/jamia/ocv130] [Medline: 26338219]
3. Hripcsak G, Albers D. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
4. Morley K, Wallace J, Denaxas S, Hunter R, Patel R, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One 2014;9(11):e110900 [FREE Full text] [doi: 10.1371/journal.pone.0110900] [Medline: 25369203]
5. Li R, Niu Y, Scott SR, Zhou C, Lan L, Liang Z, et al. Using electronic medical record data for research in a Healthcare Information and Management Systems Society (HIMSS) Analytics Electronic Medical Record Adoption Model (EMRAM) Stage 7 Hospital in Beijing: cross-sectional study. JMIR Med Inform 2021 Aug 03;9(8):e24405 [FREE Full text] [doi: 10.2196/24405] [Medline: 34342589]
6. Schleyer T, Song M, Gilbert G, Rindal B, Fellows J, Gordan VV, et al. Electronic dental record use and clinical information management patterns among practitioner-investigators in The Dental Practice-Based Research Network. J Am Dent Assoc 2013 Jan;144(1):49-58 [FREE Full text] [doi: 10.14219/jada.archive.2013.0013] [Medline: 23283926]
7. Wang S. Opportunities and challenges of clinical research in the big-data era: from RCT to BCT. J Thorac Dis 2013 Dec;5(6):721-723 [FREE Full text] [doi: 10.3978/j.issn.2072-1439.2013.06.24] [Medline: 24409345]
8. Pendergrass S, Crawford D. Using electronic health records to generate phenotypes for research. Curr Protoc Hum Genet 2019 Jan;100(1):e80 [FREE Full text] [doi: 10.1002/cphg.80] [Medline: 30516347]
9. Kim HH, Kim B, Joo S, Shin S, Cha HS, Park YR. Why do data users say health care data are difficult to use? A cross-sectional survey study. J Med Internet Res 2019 Aug 06;21(8):e14126 [FREE Full text] [doi: 10.2196/14126] [Medline: 31389335]
10. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018 Jul 20;1(1):53-68 [FREE Full text] [doi: 10.1146/annurev-biodatasci-080917-013315] [Medline: 31218278]
11. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013 Dec 01;20(e2):206-211 [FREE Full text] [doi: 10.1136/amiajnl-2013-002428] [Medline: 24302669]
12. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med 2016 Jul;71:57-61 [FREE Full text] [doi: 10.1016/j.artmed.2016.05.005] [Medline: 27506131]
13. Vezyridis P, Timmons S. Evolution of primary care databases in UK: a scientometric analysis of research output. BMJ Open 2016 Oct 11;6(10):e012785 [FREE Full text] [doi: 10.1136/bmjopen-2016-012785] [Medline: 27729352]
14. Herrett E, Gallagher A, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 2015 Jun;44(3):827-836 [FREE Full text] [doi: 10.1093/ije/dyv098] [Medline: 26050254]
15. Blak B, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care 2011 Jul 01;19(4):251-255 [FREE Full text] [doi: 10.14236/jhi.v19i4.820] [Medline: 22828580]
16. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. Inform Prim Care 2004 Feb 01;12(1):49-50. [doi: 10.14236/jhi.v12i1.108] [Medline: 15140353]
17. 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 Serv Res 2009 Sep 04;9:157 [FREE Full text] [doi: 10.1186/1472-6963-9-157] [Medline: 19732426]

18. Al Sallakh MA, Vasileiou E, Rodgers S, Lyons R, Sheikh A, Davies G. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. Eur Respir J 2017 Jun;49(6) [FREE Full text] [doi: 10.1183/13993003.00204-2017] [Medline: 28619959]
19. Lu M, Chacra W, Rabin D, Rupp LB, Trudeau S, Li J, et al. Validity of an automated algorithm using diagnosis and procedure codes to identify decompensated cirrhosis using electronic health records. CLEP 2017 Jul;9:369-376 [FREE Full text] [doi: 10.2147/clep.s136134]
20. Manuel D, Rosella LS, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. Br Med J 2010 Aug 19;341:c4226 [FREE Full text] [doi: 10.1136/bmj.c4226] [Medline: 20724404]
21. Nicholson A, Tate A, Koeling RC, Cassell JA. What does validation of cases in electronic record databases mean? The potential contribution of free text. Pharmacoepidemiol Drug Saf 2011 Mar;20(3):321-324 [FREE Full text] [doi: 10.1002/pds.2086] [Medline: 21351316]
22. Springate D, Kontopantelis E, Ashcroft D, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One 2014;9(6):e99825 [FREE Full text] [doi: 10.1371/journal.pone.0099825] [Medline: 24941260]
23. Bhattarai N, Charlton J, Rudisill C, Gulliford M. Coding, recording and incidence of different forms of coronary heart disease in primary care. PLoS One 2012;7(1):e29776 [FREE Full text] [doi: 10.1371/journal.pone.0029776] [Medline: 22276128]
24. Gulliford M, Charlton J, Ashworth M, Rudd A, Toschke A, eCRT Research Team. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PLoS One 2009 Sep 24;4(9):e7168 [FREE Full text] [doi: 10.1371/journal.pone.0007168] [Medline: 19777060]
25. Sargeant J, O'Connor AM, Dohoo I, Erb H, Cevallos M, Egger M, et al. Methods and processes of developing the Strengthening the Reporting of Observational Studies in Epidemiology - Veterinary (STROBE-Vet) statement. J Vet Intern Med 2016 Nov;30(6):1887-1895 [FREE Full text] [doi: 10.1111/jvim.14574] [Medline: 27859753]
26. Harron K, Benchimol E, Langan S. Using the RECORD guidelines to improve transparent reporting of studies based on routinely collected data. Int J Popul Data Sci 2018 Jan 10;3(1):2 [FREE Full text] [doi: 10.23889/ijpds.v3i1.419] [Medline: 30542668]
27. Benchimol E, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015 Oct;12(10):e1001885 [FREE Full text] [doi: 10.1371/journal.pmed.1001885] [Medline: 26440803]
28. Smith M, Turner K, Bond R, Kawakami T, Roos LL. The concept dictionary and glossary at MCHP: tools and techniques to support a population research data repository. Int J Popul Data Sci 2019 Dec 05;4(1):1124 [FREE Full text] [doi: 10.23889/ijpds.v4i1.1124] [Medline: 32935033]
29. McQuarrie E, Krueger R. Focus groups: a practical guide for applied research. J Mark Res 1989 Aug;26(3):371 [FREE Full text] [doi: 10.2307/3172912]
30. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
31. Onwuegbuzie A, Dickinson W, Leech N, Zoran A. A qualitative framework for collecting and analyzing data in focus group research. Int J Qual Methods 2009 Sep 01;8(3):1-21 [FREE Full text] [doi: 10.1177/160940690900800301]
32. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: Sage Publications; 2013.
33. Patel D. Research data management: a conceptual framework. Libr Rev 2016 Jul 04;65(4/5):226-241 [FREE Full text] [doi: 10.1108/lr-01-2016-0001]
34. Piwowar H, Vision T. Data reuse and the open data citation advantage. PeerJ PrePrints 2013;1:1-25 [FREE Full text] [doi: 10.7287/peerj.preprints.1v1]
35. Viseur R. Open science: practical issues in open research data. In: Proceedings of 4th International Conference on Data Management Technologies and Applications - DATA. 2015 Presented at: 4th International Conference on Data Management Technologies and Applications - DATA; 2015; Colmar, Alsace, France p. 201-206. [doi: 10.5220/0005558802010206]
36. Childs S, McLeod J, Lomas E, Cook G. Opening research data: issues and opportunities. Rec Manag J Internet 2014;24(2):142-162 [FREE Full text] [doi: 10.1108/rmj-01-2014-0005]
37. Dai S, Li H, Xiong J, Ma J, Guo H, Xiao X, et al. Assessing the extent and impact of online data sharing in eddy covariance flux research. J Geophys Res Biogeosci 2018 Jan 15;123(1):129-137 [FREE Full text] [doi: 10.1002/2017jg004277]
38. de Almeida UB, Fraga B, Giommi P, Sahakyan N, Gasparyan S, Brandt C. Long-term multi-band and polarimetric view of Mkn 421: motivations for an integrated open-data platform for blazar optical polarimetry. Galaxies 2017 Nov 30;5(4):90 [FREE Full text] [doi: 10.3390/galaxies5040090]
39. Cragin M, Palmer C, Carlson J, Witt M. Data sharing, small science and institutional repositories. Philos Trans A Math Phys Eng Sci 2010 Sep 13;368(1926):4023-4038 [FREE Full text] [doi: 10.1098/rsta.2010.0165] [Medline: 20679120]
40. Ceci SJ. Scientists' attitudes toward data sharing. Sci Technol Hum Val 2018 Mar 01;13(1-2):45-52 [FREE Full text] [doi: 10.1177/016224398801300106]

6 CHAPTER 6: RESULTS 1 – CONCEPT LIBRARIES FOR REPEATABLE AND REUSABLE RESEARCH: QUALITATIVE STUDY EXPLORING THE NEEDS OF USERS

6.1 INTERVIEWS WITH USERS

In total, 6 one-to-one interviews were conducted, and each interview lasted for approximately half an hour. The analysis of the interviews resulted in 4 main themes, with several subthemes. The four main themes are as follows:

1. Previous opinion of a prototype concept library
2. Requirements of a prototype concept library
3. Experience of existing concept libraries
4. Recommendations to improve repeatable research

Table 6.1 shows the main themes and subthemes of one-on-one interviews, along with examples of participant narratives.

Table 6.1: Presentation of the themes and subthemes of one-to-one interviews

Themes	**Examples of participant narratives**
Theme (1): Prior opinion of a prototype concept library	
Positive	*"If there's a way of doing that already that is set up and is validated and is consistently applied that would be an amazingly useful resource"* (researcher 2).
Neutral	*"It will be helpful, but it needs to be extended. If they want to build something like this, and it is effectively working as a library, you need two things to be happened: 1) people are happy to feed in their constructs so it builds up, and 2) a useful library, easy to go, to browse, and to borrow phenotypes definitions"* (a clinician).
Negative	None

Theme (2): Requirements of a prototype concept library	
Usability Simplicity	*"Simple plain English not in SQL or python"* (a clinician).
Searching ability	*"What is the type of search engine? Is it a search engine that just does disease phenotypes or also does the health status phenotypes or risk factor phenotypes, symptoms phenotypes"* (a clinician)?
Data Quality	*"It's really just about transparency and documentation. So, anybody can effectively do anything that can be turned into a reproducible research output. The barriers are usually not enough time to comment and document it properly and then not enough quality assurance"* (a senior research manager).
Sharing ability	*"It would be very useful to share the knowledge about codes such as Read Codes, ICD 10 codes, or OPCS codes, and share ideas and concepts between other users that will save lots of time"* (researcher 3).
Sustainability Interoperability	*"How interoperable it is with other systems because the major failure of most of these systems is that they're not interoperable, so people don't use them"* (a senior research manager).
Accessibility	*"So, from a group like myself, or me as a user, we would probably like direct access to the underlying data it stores. So, whether that's through something like SQL directly, or something like that through a statistical package, because where we do lots of bulk type work"* (a senior research manager).
Analysability	*"I wanted to look at all health codes of my study population. Then, through machine learning, like feature selection, I tried to identify the most important list of codes, which are associated with the popular health conditions"* (researcher 1).
Theme (3): User experience of existing concept libraries	

Aware-used them	*"Yes, so with QOF, we definitely used QOF codes a lot, because obviously going back to the quality assurance question, they'd been assured so that the NHS can use them for remuneration of money and payments. With other systems, we tend to look online to see CALIBER of things with us, then yes, we have used outputs from those systems before"* (a senior research manager).
Aware-not used them	*"No. I have not used any of these things before so I think there is CALIBER and I think, is that part of what was set up within the previous Farr institute? so I am aware that some of these exist but I haven't looked into them before"* (researcher 2).
Not aware	None
Theme (4): User's recommendation to improve repeatable research	*"If we want reproducible research, we have to all be using these resources in a similar way or at least we need to be able to understand what previous projects have done. It is about setting things out clearly. Clear definitions, clear sets of codes that people can then either use themselves or build on I think"* (researcher 2).

6.1.1 Previous opinion of a prototype concept library

The majority of the participants were positive about the prototype concept library and felt that a concept library in principle could be a very helpful resource for conducting repeatable research. A machine learning expert mentioned that a concept library will be an extremely useful resource because Read codes from general practice and International Classification of Diseases (ICD)–10 codes from hospitals are the most common data items that machine learning experts would like to use most often. They use data linkage repositories to extract the necessary data for machine learning in public health studies, and they use the codes to extract the data from the repositories:

Researcher 3 said, *"It would be very useful to share the knowledge about codes such as Read codes, ICD 10 codes, or OPCS codes, and share ideas and concepts between other users that will save lots of time. It is useful to use verified codes,"* and researcher 2 stated, *"If there's a way of doing that. Already that is set up, and is validated and is consistently applied, that would be an amazingly useful resource."*

However, 2 participants (a clinician and a senior research manager in health data science) were not sure about the effectiveness of the prototype concept library because they felt that users had to engage with it for it to be useful and they were not sure how well users would engage:

"There is potential that it could be useful as a tool. It will kind of come down to how usable it is, how flexible it is, how well it's maintained, how much of the community uses it" (a senior research manager).

6.1.2 Requirements of the prototype concept library

The participants mentioned several requirements they would like to see in the prototype concept library. For example, they stated that the concept library needed to have high usability. This means that it needs to be simple and easy to use by naïve users:

"It should be simple enough, within one or two clicks; we can find the required data, but also should contain advanced expert features (R, SQL, or Python programming languages) to extract, include, or exclude codes necessary for their studies" (researcher 3) and *"Like, in one of my previous projects, I looked at, from a machine learning perspective, I wanted to look at all health codes of my study population. Then, through machine learning, like feature selection, I tried to identify the most important list of codes, which are associated with the popular health conditions"* (researcher 1). They also stated that the concept library should have a good search engine so that they can easily find the phenotypes and phenotyping algorithms they want to use. A clinician inquired, *"What is the type of search engine you are developing? Is it a search engine that just does disease phenotypes? or also health status phenotypes, risk factor phenotypes, or symptom phenotypes. For example, I am looking for diabetes, but I may also be looking for smoking or alcohol consumption, or symptoms like pain or cough. So, how big is the enterprise and how do you search for what are the appropriate terms? Discussion is needed to know what is it?"*

In addition, the participants stated the following requirements:

1. Include the data sources used (e.g., codes from general practice, hospital [ICD and Systematized Nomenclature of Medicine], and British National Formulary medication), a general clinical code list for comparison, lists of ontologies along with their variances and versions, and a description of how codes were established:

   "It is about setting things out clearly. Clear definitions, clear sets of codes that people can then either use themselves or build on, I think" (researcher 2).

2. Have a clear phenotyping algorithm labelling convention for search engines:

A clinician stated, *"What do you search on? Thought about what do you call these phenotypes? Is there a consistent in calling them? For example, Type II diabetes, or insulin dependent diabetes"* and researcher 1 stated, *"So, first of all, for the code reference library, two things are always there in my mind. It's in my opinion again. Number one, they should be validated. Secondly, they should be correctly labelled."*

3. Specify why a particular phenotyping algorithm was developed (e.g., definite disease or probable/suspected condition definitions):

   "When I have an algorithm, I want a field that tells me the purpose of the algorithm, a brief description of what the algorithm is intended to do" (a clinician).

4. Illustrate the logic model category used to create phenotyping algorithms (i.e., code lists, inclusion or exclusion factors, and clinical or machine learning approach used):

   "Is this just a code list of inclusion factors? And or exclusion factors? Or is it static? Does it have a tampered relationship? So, some algorithms are present or absence of conditions, some required a tampered dependence. In the logic model categories: Is this a clinically derived algorithm from experts' views or for instance that machine learning derived algorithms" (a clinician).

5. Use ready-made phenotyping algorithms that can be modified to fit the needs of their research. All participants agreed that if they had to create their own phenotyping algorithms because ready-made phenotyping algorithms could not be modified, they needed an easy approach to use a code list in the concept library.

There was an issue regarding how to validate phenotyping algorithms, and most participants expressed their preferences for using all possible methods of identifying a condition, valid or not, to allow the researcher to choose the phenotyping algorithms according to their research requirements:

"So, there is no right answer for that because it's going to be very dependent on your research question, your study group, and your study design. So, once again, if the concept tool is going to match multiple different use cases, it's going to need to accommodate for those different types of study design" (a senior research manager).

Sharing phenotyping algorithms needed to be easy and not time-consuming, and some felt there needed to be some recognition of their work before they would give their codes. Finally, a concept library must be interoperable with other products or systems:

"How interoperable it is with other systems, because the major failure of most of these systems is that they're not interoperable, so people don't use them" (a senior research manager).

Most participants wanted the source code (e.g., the SQL code for the phenotyping algorithm itself) to be available in a downloadable machine-readable format to be able to access it using specific programming languages such as R, SQL, or Python.

6.1.3 Experience of existing concept libraries

All participants stated that they were aware of some existing concept libraries, such as CALIBER and ClinicalCodes.Org (both in the United Kingdom), but most of them did not use them. The reasons given for not using them were that they already had their own self-made concept libraries (e.g., concepts they have used before) or the available concept libraries did not provide phenotyping algorithms that fit their studies. For example, a machine learning expert mentioned the reasons for not using two of the existing concept libraries, namely the Concept Dictionary at the Manitoba Centre for Health Policy in Canada and CALIBER in the United Kingdom were,

"Canadian systems provide Canadian data for their studies; CALIBER is specific for cardiovascular disease and does not have many concepts in it."

Conversely, 2 of the participants mentioned that they used some existing concept libraries to extract and develop phenotyping algorithms for their studies:

"We definitely used QOF codes a lot, with other systems, we tend to look online to see CALIBER, we have used outputs from those systems before" (a senior research manager).

6.1.4 Recommendations to improve repeatable research

The participants suggested the following recommendations to improve repeatable research output using routine data:

1. There should be a drive for more transparency in research methods documentation, such as publishing complete phenotype definitions and clear code lists:

   A senior research manager stated, *"It's really just about transparency and documentation. So, anybody can effectively do anything that can be turned into a reproducible research output,"* and researcher 2 said, *"If we want reproducible research, we have to all be using these resources in a similar way*

or at least we need to be able to understand what previous projects have done. It is about setting things out clearly. Clear definitions, clear sets of codes that people can then either use themselves or build on, I think."

2. Providing opportunities for researchers to collaborate rather than working in isolation:

   "The barriers are usually not enough time to comment and document it properly and then not enough quality assurance. So, if there was more time and or more availability of those kinds of opportunities for people to collaborate rather than doing things in isolation, there's almost all the research we do here could be turned into a reproducible type of output" (a senior research manager).

3. Develop a concept library that enables researchers to begin classifying population outcomes using uniform codes:

   "I think that a resource like this is a very good step in the right direction because I think what people need to start doing is using consistent codes in order to identify conditions or outcomes within populations" (researcher 2).

4. Provide validated phenotyping algorithms that researchers can use directly to avoid duplication, with the ability to modify them to meet their own research needs:

   "For each project, it always has some specific requirement which is unique, which is not common. There are some things which are common, and there are a few things which are very unique. So, we need to have some algorithms which we can just use to, you know, just to avoid the duplication, but also, we need to have control of the algorithms, so that we know only that these bits are going to be different for this project, so I'm going to replace, change, modify this bit, and we'll run it" (researcher 1).

6.2 FOCUS GROUP DISCUSSIONS

Of the 34 invited researchers, 14 (41%) attended the focus group discussion. These participants were researchers (14/34, 41%) from Swansea University who were working with the SAIL data in the Data Science Building. Of the 14 participants, 5 (36%) were female participants and 9 (64%) were male participants. Furthermore, 6 (43%) participants were PhD holders, 6 (43%) were Master's degree holders, and 2 (14%) were Bachelor's degree holders.

The focus group discussion was held for 2 hours to perform a SWOT analysis of the current system for phenotyping and the proposed concept library, which was recorded and transcribed. Figure 6.2 shows the SWOT Analysis Framework.

Figure 6.2: SWOT analysis framework

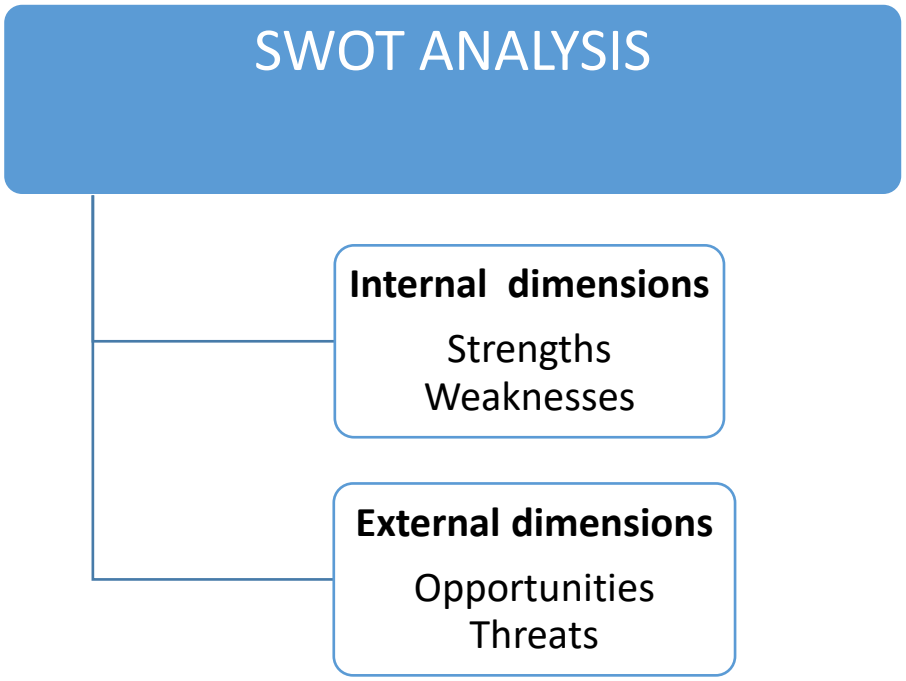


Table 6.2 presents a summary of the SWOT analysis of the current system for phenotyping and the proposed concept library. The 2 focus group discussions were audio recorded and transcribed verbatim. Thematic analyses were then conducted using the 6 steps of Braun and Clarke to discover the main themes and subthemes [118].

Table 6.2 A summary of a SWOT analysis of the current system for phenotyping and the prototype concept library

SWOT Analysis
STRENGTHS • Concept libraries provide researchers with a good starting point. • Publicly available code lists may provide researchers with a history of a particular area of research, such as asthma. • Referencing previously published lists of codes enables researchers to demonstrate a rationale for using such lists of codes. • Using research methods developed by others that match the researchers' interests could result in significant time savings. • Collaboration among researchers is facilitated through sharing and using research methods such as code lists.
WEAKNESSES • Searching for and reusing phenotypes and codes is a time-consuming and labour-intensive process. • There are various lists of codes for each phenotype definition. • The list of codes chosen by clinicians varies significantly. • A large number of previously developed code lists could not be repeated. • Reusing other researchers' data requires programming knowledge such as SQL. • Some of the ready-made phenotyping algorithms may not be very useful in terms of their general purpose. • Some existing concept libraries have limited user interfaces. • Some existing concept libraries are not user-friendly. • It is unclear who is accountable for the quality of the uploaded codes in concept libraries. • The validity of the content of concept libraries is unclear.

OPPORTUNITIES

- Concept libraries must provide user documentation.
- Concept libraries must provide users with training.
- Transparency in sharing the whole approach used to create the code lists is required.
- Establishing a standardised way of defining each specific condition in order to facilitate comparisons of research outcomes across the United Kingdom.
- Creating a specialised library that stores code lists of a specific condition within a specific set of patients, such as a concept library specialising in chronic conditions in children.
- Creating a concept library that engages a wide variety of users (i.e., is easily understandable by clinicians but has some advanced features such as programming skills for more expert users).

THREATS

- The inconsistency of data across various databases makes data reuse difficult.
- Lack of confidence in the quality of the list of codes developed by other researchers if they are not cited.
- Access to code lists is limited since some researchers do not publish them alongside their studies.
- Different research outcomes result from a lack of access to a list of codes created by other researchers.
- Data sharing may be inhibited if there are no returns, such as referencing and acknowledgement.
- Concerns about ownership rights discourage data sharing (for example, methods could be used as their own by other researchers before publication).

Thematic analysis was conducted on the transcripts, which resulted in the identification of the following seven main themes:

1. Facilitators for and barriers to participants' contributing their research methods
2. Facilitators for and barriers to participants' use of other researchers' methods

3. Participants' concerns about the prototype concept library
4. The requirements of the participants for the prototype concept library
5. Participants' recommendations to improve repeatable research
6. Participants' perceptions of their current phenotyping system
7. Participants' use and perceptions of existing concept libraries

6.2.1 Facilitators for and barriers to participants' contributing their research methods

6.2.1.1 Facilitators

Several facilitators were identified by participants as motivators for them to share their work (e.g., phenotyping algorithms and code lists). Many participants stated that being credited appropriately (e.g., receiving citations from other researchers) would motivate them to share their work:

"If whoever's using it acknowledges it's use in whatever they publish, at least you're getting some recognition" (data scientist 8) and *"If there were DOIs attached to the code list of algorithms, when people are publishing, there's an incentive for putting it on there, because they're able to demonstrate the impact their work has had"* (data scientist 4).

Some participants stated that communicating with their research team would encourage them to organize team resources and discuss research findings from other researchers who used their code lists. However, improving research opportunities, increasing academic achievement, and sharing knowledge through collaboration with other researchers working in the same organization would motivate some of the participants to share their work:

"I think there's benefit to the organization, and there has to be benefit to the people contributing to it" (data scientist 4).

In general, researchers work in an organization (e.g., a university or a research institute), and they work hard to improve the research outcomes of their organization. Some participants stated that advancing the research base and saving other researchers' time and effort would stimulate them to share their work:

"Surely if you've done something you think really worthwhile, you want other people to use it, as well, because then that furthers the research" (data scientist 6).

6.2.1.2 Barriers

On the other hand, the participants pointed out several barriers that could inhibit them from sharing their work (e.g., phenotyping algorithms and code lists) with other researchers. Some participants argued that it is easy to build a phenotyping algorithm that fits exactly their needs, but it is more challenging to develop a general one, so it can be used by others (e.g., many clinical researchers have created phenotyping algorithms for particular research, and these algorithms are difficult to generalize).

Several participants mentioned that a lack of return for their hard work (e.g., not receiving any credit from others, such as referencing when they reuse their data) would prevent them from sharing their work:

"How do you enforce that people are going to give you credit? It doesn't happen sometimes, when referencing, saying where they got it from. You've just got to hope they do" (data scientist 11).

Some participants were worried about their intellectual rights (e.g., if they shared their methods such as phenotyping algorithms before publication, other researchers would use them as their own).

6.2.2 Facilitators for and barriers to participants' use of other researchers' research methods

6.2.2.1 Facilitators

The participants mentioned several facilitators that would encourage them to reuse research methods developed by others, such as the following:

1. Using existing code lists can save them a lot of time and effort, which they frequently spend creating new code lists from scratch:

   "It's the first stage of every single process, and we tend to get two or three months of work, until we get to that final code list, and we can now start looking at the cases" (data scientist 10).

2. Reusing available data, such as code lists, is a good place to start for researchers (for example, they can use them to examine new ideas and gain new insights):

"Having code lists would be such a help, to get you started. They always want things like BMI and weight and height. There are hundreds of codes for those. The smoking codes, having a list, even if you don't use the algorithm that they've developed, is a huge bonus" (data scientist 12).

3. Using the work of others as a reference to compare research outcomes, and researchers want to prove that there is a basis for the use of such codes.

6.2.2.2 Barriers

Conversely, the participants pointed out several barriers that could inhibit them from reusing methods developed by other researchers such as the following:

1. Poor data quality discourages researchers from reusing it:

   "You could upload complete garbage" (data scientist 1).

2. Some phenotyping algorithms will not work outside the population in which they were developed. For example, code developed in Canada may not be relevant to finding conditions in general practitioner data in the United Kingdom:

   "Yes, it works in their population, because where they've trained it." (data scientist 5).

3. Whether the data are useful to researchers plays an important role in the decision to reuse them (e.g., researchers would not use a phenotyping algorithm if its general purpose did not match their interests):

   "Yes, a general-purpose algorithm may or may not be very useful to have it to see what they've done, but you may not use it" (data scientist 12).

6.2.3 Participants' concerns about the prototype concept library

When researchers decide where to deposit, share, and reuse data, they prefer to use approved concept libraries: *"Is it going to be approved?"* (a financial planner). Moreover, some participants stated that it is not clear who is responsible for the quality of the phenotyping algorithm, if this is the responsibility of the developers running the concept library or the responsibility of the researchers uploading the phenotyping algorithms:

"If people send the codes, the onus of the quality of that code list you would still want to be the responsibility of the researcher to be submitting worthwhile codes. You don't

want to then be the guardian of the quality of the code list. You still need to know where the responsibilities lie" (data scientist 4).

Researchers do not want to upload phenotyping algorithms if they could be blamed for flaws, and health informatic developers do not want to take responsibility for the phenotyping algorithms that were uploaded.

The participants expressed their concerns about the completeness rate of the phenotyping algorithms. They would like to know the percentage of the gap to be considered when using a phenotyping algorithm from the prototype concept library:

"What is the completeness rate? For certain things, we know there are gaps. If the gap is 20%, is that something I should be including in any algorithm I'm considering?" (data scientist 8).

In addition, there has been a question as to whether codes need to be peer reviewed so that quality is evaluated.

6.2.4 Requirements of the participants for the prototype concept library

6.2.4.1 Usability

1. Learnability: Some participants said they would like the concept library to be easily understandable by clinicians, who acknowledge the clinical definition of the code lists with little technical skills to simply point and click the selected code lists, whereas other participants requested the availability of advanced functions to be used by expert users:

   "The concept library should be easy. Someone needs to train us" (data scientist 9).

2. User documentation: A collection of well-defined task-oriented documentation for users was required by some participants. They want a user documentation that consists of clear, step-by-step instructions on how to use the concept library and gives examples of what the user can see at each step (e.g., screenshots would be useful):

   "Concept library should have some documentation" (data scientist 9).

3. Data quality: Some participants required the availability of a consistent method for identifying each specific condition to ensure that what researchers are doing

is compatible within their immediate team but also within the broader research community in the United Kingdom to facilitate a comparison of research outcomes. Other participants stated that they needed a predefined list and a uniform approach describing how to use existing codes of additional diagnoses, such as smoking:

"Additional things like smoking and alcohol status are used a lot, but they're usually very different for every project. We should have a more uniform way of doing it, like, we'll take that bit off the shelf and use it, and do the bespoke bit for things that need to be bespoke" (data scientist 5).

If there are multiple code lists for the same condition, some participants proposed that versions be generated to describe each particular condition:

"*So, it would be relevant that there were multiple lists for the same condition, if you've got a version and way of defining a certain condition"* (data scientist 4).

4. Transparency: Several participants required transparency in sharing the entire approach used in developing the code lists, including phenotyping algorithms and the methods used. They stated that if they use a code list for each comorbidity of a condition, they will build an entirely different score over the years. Therefore, transparency in the documentation of research methods would help them to know which score is the best.

6.2.4.2 Sustainability

1. Accessibility: Several participants needed the availability of an access control that allows access to the codes only after publication, while at the first stage of the study, researchers spent a lot of time and effort developing them, and they feared someone else could publish work faster than them using the algorithms:

   "There should be an option in the concept library for lists that have been published. People can develop them, but if they're not published, you don't have to use them" (data scientist 3).

2. Licensing: Some participants needed to know which type of license was adopted by the developers of the concept library (e.g., researchers can have one that means any researcher can take it and use it, or they can have one that means researchers can use it but not for commercial purposes).

3. User community: Several participants required users to quote a reference if publishing papers based on the results (partially or completely) derived from the concept library:

   "If I want to use someone else's work, I think that's the norm, and should be in this economy. Anything, not just code. To use this, I should reference that it's based on this or other thing completely, or a part of it" (data scientist 2).

   Referencing helps to determine whether there is or will be an active user community for the concept library and the codes used:

   "It potentially would make your publication more discoverable. If there's a whole community of users using this" (data scientist 1).

6.2.5 Participants' recommendations to improve repeatable research

Of the 14 participants, 9 (64%) suggested that the prototype concept library should be accessible both in the United Kingdom and globally and practically available to enable researchers around the world to use a web-based secure platform, which stores codes and other logic, and to encourage researchers to contribute their codes to promote research:

"Should be open for the United Kingdom" (data scientist 9).

However, a participant recommended that the prototype concept library should be closed at the beginning to ensure it is working and then to become opened as researchers build trust:

"You might need to restrict it, to start with, to make sure it works. Otherwise, everyone will see the problems you might have" (data scientist 12).

In addition to know who is using the concept library, data scientist 8 suggested that it should have request sharing followed by open sharing. Accessibility to research data has significant potential for scientific advancement as it promotes the replication of research results and enables the use of old data in new contexts. With respect to this, some participants suggested that funders and publishers should obligate researchers to share their research data such as code lists:

"Some sort of obligation by funders to share this" (data scientist 2) *and "Publishers, as well"* (data scientist 8).

A participant suggested the use of preauthorization of publication by journals based on the research protocol because researchers can put their protocol first, and all the limitations are actually corrected before they run the research. This approach has many advantages for both the researcher and the publisher, as it improves the quality of the output. Another participant recommended the creation of a discussion forum in the concept library to facilitate collaboration among researchers on just about any topic (e.g., they can share their ideas, submit their comments, and discover new ideas):

"Make it almost a forum" (data scientist 8).

6.2.6 Participants' perceptions of their current phenotyping system

The participants mentioned several problems associated with the current phenotyping system. For example, they have to search for codes from different databases, which use different coding systems such as Read codes and ICD-10 codes, and then they have to validate the selected code lists with experts in the field such as clinicians:

"I have to google all of this and search what was there within the community. I have to go to CALIBER, I have to go to Manchester, or there is a work in Edinburgh University, do some work there. Do the search. I have to go there, see the ability to work, and start. It does take a lot of time. Based on my study of Google, I have to start a record, and I have to validate it, verify with other people, clinicians or researchers. It's a long process" (data scientist 9).

Although they could find some codes on the web, they still had to locate the list manually, copy it, and enter the codes into their scripts. Often, they might spend a few days on it, and they might miss obscure codes or even use irrelevant codes:

"Starting from scratch, I would go online to see what's available. Go into other people's and see their code lists" (data scientist 11).

With respect to this, some participants said that they preferred to use code lists that were referenced or used by other researchers. Some participants reported that the Read code lists chosen by the researchers were different from the Read code lists chosen by general practitioners. For example, they found that there were some very clear codes, but they were rarely used by general practitioners:

"What we get in the Read code list isn't necessarily what the GPs are recording it under" (data scientist 12).

They also stated that there is a significant difference between what one general practitioner may say in a list of codes versus another:

"*For example, there is no single entity code for asthma. There are different entities. If you want to find specific things within asthma, there's a list of codes for them"* (data scientist 2).

6.2.7 Participants' use and perceptions of existing concept libraries

Not all participants had previously used some of the existing concept libraries. However, most of those who used some of them expressed negative perceptions. For example, several participants stated that the concept libraries they used were not user-friendly (i.e., they were difficult to use by new users):

"For CALIBER, it seems not so user friendly. It's not easy. You have to know first. Someone needs to train you up. For new users, it's difficult to get inside CALIBER. The concept library should be easy. Someone needs to train us. Concept library should have some documentation" (data scientist 9).

Therefore, training and good user documentation are required. A further problem for some participants was the inconsistency of data among various databases, which makes reuse of data quite challenging:

"But if there is something that gets secondary and primary care involved, and there's a registry, if the definitions that are created in Manchester, how easy will it be to apply it to, for example, in Wales or Scotland, where registry is a bit different?" (data scientist 8).

Participants who did not use any of the existing concept libraries expressed different perceptions about them. For example, some participants reported that they wanted to explore available concept libraries. Others, however, expressed doubts about the quality and validity of the data stored in these concept libraries, which could prevent them from using them:

"I haven't looked at them myself, but if you go on this clinical code site and you type in diabetes, there are 50 different code lists people have put together for diabetes" (data scientist 6).

Some participants stated that the main reason for not using any of the existing concept libraries is not finding a concept library that matches their studies. The developers of concept libraries may consider building a specialized library that stores code lists of a particular condition within a specific group of patients according to researchers' needs, such as developing a concept library that specializes in chronic conditions in children.

Chapter 7

Results 2 - Quantitative study

This chapter presents the results and statistics of the quantitative study that was performed using the Concept Library Usability Scale, an e-mail survey instrument. The purpose of this study was to investigate the opinions of users from various disciplines, such as academics, clinicians, machine learning experts, and research managers about the usability of concept libraries.

7 CHAPTER 7: RESULTS 2 – USING THE CONCEPT LIBRARY USABILITY SCALE TO INVESTIGATE THE USABILITY OF CONCEPT LIBRARIES

7.1 OVERVIEW OF THE QUANTITATIVE STUDY

This chapter presents the results and statistics of the quantitative study that was performed using the Concept Library Usability Scale, an e-mail survey instrument that was developed to investigate the opinions of users from various disciplines, such as academics, clinicians, machine learning experts, and research managers about the usability of concept libraries (See figure 3.7 in chapter 3). The e-mail survey contains 12 statements, and participants choose one of five response options ranging from "strongly disagree" to "strongly agree" to express their level of agreement or disagreement with the first ten statements. The eleventh statement was intended to allow participants to openly express and discuss their thoughts, while the twelfth statement invited participants to submit contact information if they were interested in participating in a one-to-one interview.

Email invitations were sent to the participants (N = 200) twice, who included researchers, health professionals, and clinicians, to explore their attitudes and opinions about the usability of one of the existing concept libraries in the United Kingdom (i.e., the CALIBER research platform), which contains "research ready" variables derived from linked EHRs from primary care, coded hospital records, social deprivation data, and cause-specific mortality data. In addition to the invitation emails, we sent out a tweet with a hashtag asking CPRD and SAIL users to complete the email survey: (https://twitter.com/SineadBr/status/1417375178900770816?s=20&t=o8jdGHQKivrzWbgBLjr4jA).

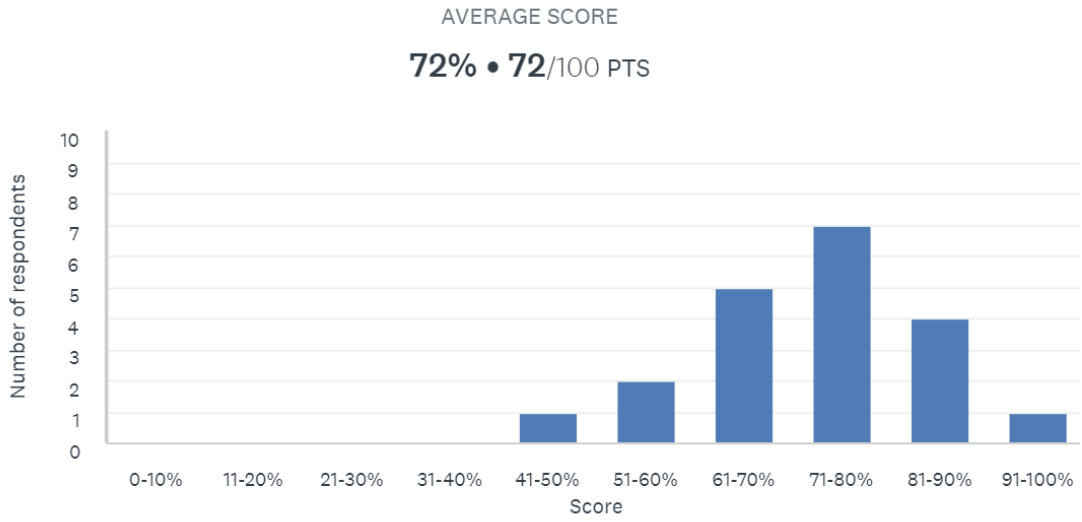
Participants were instructed in the e-mail invitations to spend approximately fifteen minutes searching the CALIBER research platform for algorithms that they would like to use for future research or that they had recently used for their studies, and then provide feedback on the resource's usefulness in this case by completing our e-mail survey, the Concept Library Usability Scale. Only 20 of the 200 invited participants completed the e-mail survey (See table 7.1).

Table 7.1 Number of invited participants, number of completed e-mail surveys, and response rate

Number of invited participants	**Number of completed e-mail surveys**	**Response rate (%)**	**Note**
200	20	10%	16 out of the 20 participants skipped Statement number 12

As mentioned in chapter 3, I used the SurveyMonkey tools for the analysis process of the results. In the design section of the Concept Library Usability Scale, the five Likert Scale, which is a variation of the Matrix statement, was used, and weights were assigned to each answer choice. Then, the Likert Scale automatically calculated a weighted average for each answer choice in the analyse results section of SurveyMonkey. Figure 7.1 shows the average score and the number of participants and their associated scores.

Figure 7.1 The average score and the number of participants and their scores



7.2 RESPONSES OF PARTICIPANTS TO THE FIRST STATEMENT: "I THINK THAT I WOULD LIKE TO USE THIS CONCEPT LIBRARY FREQUENTLY"

The purpose of the first statement was to identify the percentage of participants who would prefer to frequently use the CALIBER research platform. The average score of this statement was 69%. According to the results in Figure 7.2, two participants (10.00%) strongly disagreed with this statement for the following specified reasons:

- *"Ontologies have to be standardized for multiple reasons including reproducibility, reliability, and comparability"* (participant 1).
- *"It is not clear who made the code lists, who did the QA, with the exact purpose, whether both persons are native speakers, whether they are medically and methodologically sufficiently qualified and consultants etc. etc"* (participant 10).

Additionally, two participants (10.00%) disagreed with the frequent use of the CALIBER research platform. They stated the following:

- *"I would use it, but not frequently"* (participant 7).
- *"Many may be outdated and not context specific. Also limited to Read v2"* (participant 12).

The results indicated that four participants (20.00 %) were neither agreed nor disagreed to frequent use of the CALIBER research platform for the following reasons:

- *"It doesn't look like it has a lot of concepts so might use it but not sure it has that many conditions covered"* (participant 2).
- *"Conditions of interest not included in library - neurodevelopmental disorders"* (participant 6).
- *"I could see myself using the library, it would likely not be that frequently. Possibly at the start of projects when exploring unfamiliar concepts" (participant 15).*
- *"I do usually check CALIBER before developing a new code list. I also check LSHTM's data compass and our electronic health records group research group, and check for validated or previously used lists with a PubMed search for studies on the topic using electronic health records. I don't find I use CALIBER frequently, but I find it a very valuable resource when I need it"* (participant 18).

On the other hand, the results showed that nine (45.00%) of the participants agreed they would use the CALIBER research platform on a frequent basis. They stated the following:

- *"I need to use code lists for my work very frequently and I trust CALIBER"* (participant 3).

- *"It is a useful source of a wide range of definitions"* (participant 4).
- *"Collection of codes that identify likely diagnosis of a condition with validation through linked publication is really useful. Saves a lot of time identifying codes, only limitation would be related to date of publication and likelihood of nest / additional codes being identified or codes being deprecated"* (participant 5).

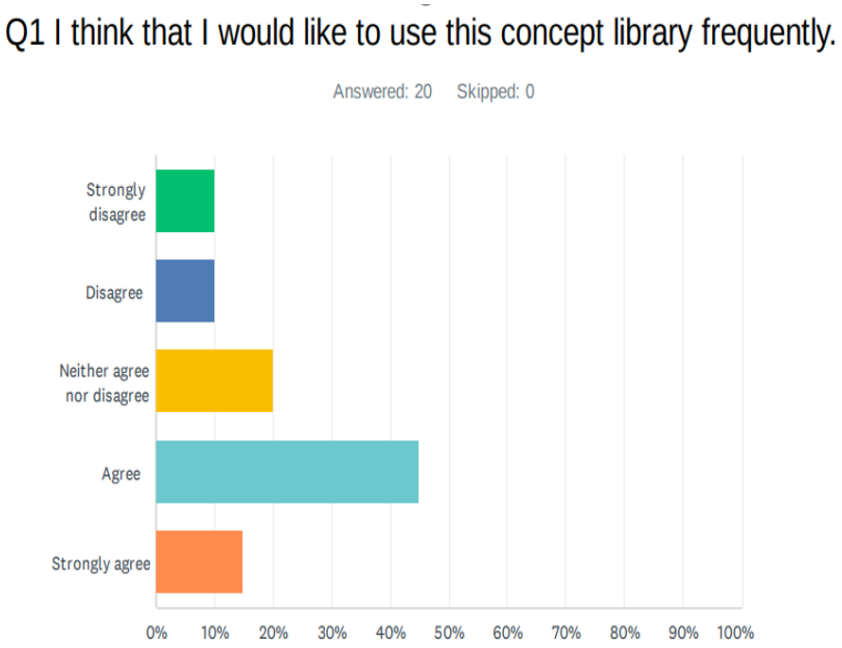
However, some of the participants agreed partially with the first statement. They mentioned that they would like to use the CALIBER research platform but not on a frequent basis. They stated the following:

- *"I'm not sure if I'd use it frequently... but I would use it to check completeness of existing codes that OPRI already has, particularly when initiating a new study. The codes, as far as I can see (I looked at half a dozen categories) don't have SNOMED in them. This will obviously be an issue going forward"* (participant 13).
- *"Useful although only has Read 2 without EMIS only codes, so two additional bits of work for EMIS data or CTV3 (we are going to add some CTV3 mapping)"* (participant 17).
- *"I agree that I would use this library, although less frequently in the future than now because of the retirement of Read codes and move to SNOMED CT codes in the CPRD (my main data source)"* (participant 19).

Three participants (15.00%) strongly agreed with the first statement:

- *"It's been invaluable in my current research project, saves reinventing the wheel each time we explore a new disease group to have the codes already defined"* (participant 8).
- *"Consistency across different studies"* (participant 9).
- *"Help facilitate the confirmation of codes and concepts"* (participant 14).

Figure 7.2 Percentages of the participants responses to the first statement



7.3 RESPONSES OF PARTICIPANTS TO THE SECOND STATEMENT: "I THINK THAT I WOULD NEED THE SUPPORT OF A TECHNICAL PERSON TO BE ABLE TO USE THIS CONCEPT LIBRARY"

The second statement was designed to identify the percentage of participants who would require the support of a technical person in order to be able to use the CALIBER research platform. The average score of this statement was (79%). Figure 7.3 showed that five participants (25.00 %) strongly disagreed with this statement for a variety of reasons, including:

- "*I am familiar with the use of Read codes and some of the other coding systems and with building scripts to use them"* (participant 4).
- *"I am a technical person"* (participant 7).
- *"Fairly straightforward"* (participant 9).

In addition, eleven participants (55.00 %) disagreed with this statement for a variety of reasons, including:

- *"Looks straightforward to use"* (participant 2).
- *"No need"* (participant 8).
- *"I can use these codes and incorporate into central SQL/local STATA formats"* (participant 13).
- *"The portal is quite easy for use"* (participant 14).
- *"Pretty straightforward"* (participant 17).

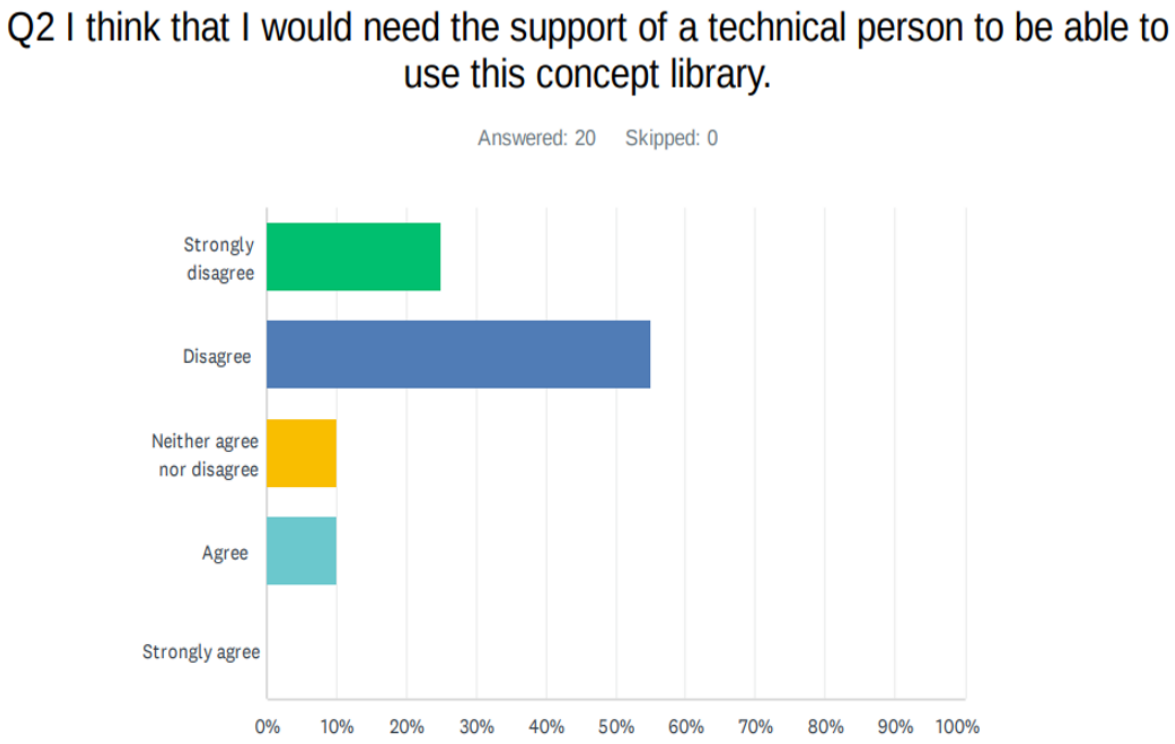
- *"It's very user friendly"* (participant 18).
- *"I have a lot of experience of collating medical code lists and using them in data manipulation and analysis using CPRD and secondary data care sources"* (participant 19).

Results showed that two participants (10.00%) neither agreed nor disagreed with the second statement. Participant 5 stated, *"Depends if you can scrape from the webpage to a csv format within SAIL"*. However, two participants (10.00 %) expressed agreement with this statement, as follows:

- *"Getting the code lists into R is difficult (the package is a nightmare, sorry). I've used copy-paste in R to make this easier"* (participant 3).

- *"I think I would like the support of an experienced user. I could learn to navigate it alone, but even instructional/explanatory videos could be useful"* (participant 15).

Figure 7.3 Percentages of the participants responses to the second statement

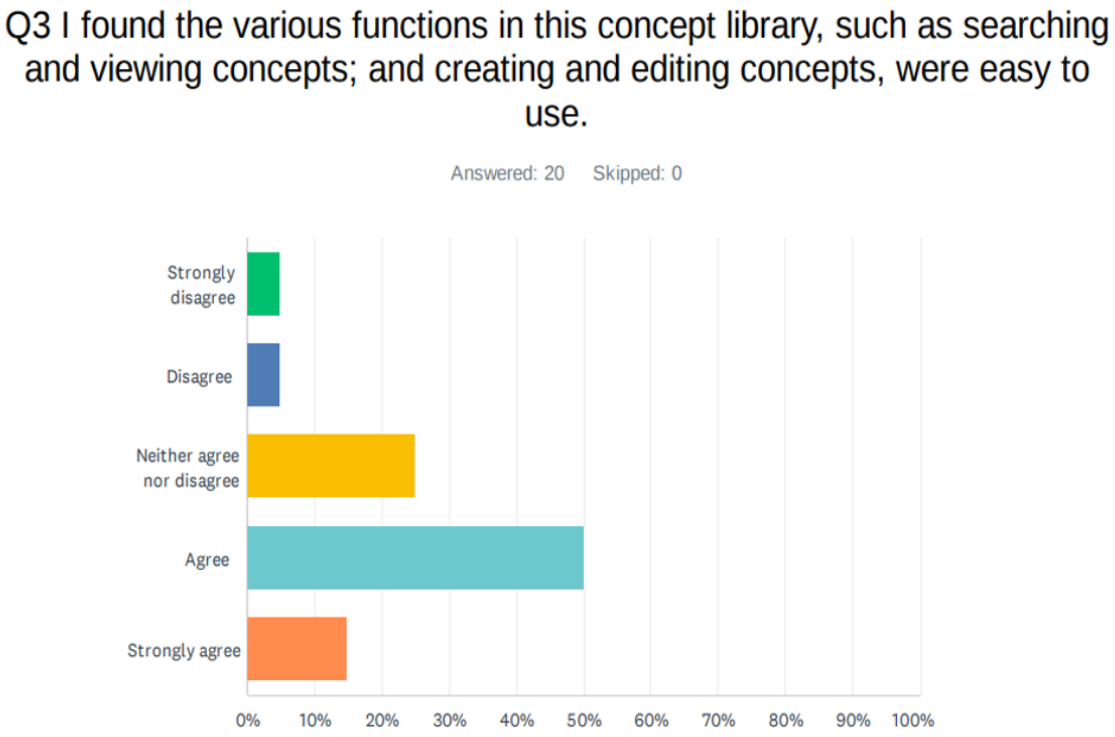


7.4 RESPONSES OF PARTICIPANTS TO THE THIRD STATEMENT: "I FOUND THE VARIOUS FUNCTIONS IN THIS CONCEPT LIBRARY, SUCH AS SEARCHING AND VIEWING CONCEPTS; AND CREATING AND EDITING CONCEPTS, WERE EASY TO USE"

The purpose of the third statement was to determine the percentage of the participants who found the various functions in the CALIBER research platform, such as searching

and viewing concepts and creating and editing concepts, easy to use. The average score of this statement was 73%. The results in figure 7.4 showed that one participant (5.00%) strongly disagreed, one participant (5.00%) disagreed, five participants (25.00%) neither agreed nor disagreed, ten participants (50.00%) agreed, and three participants (15.00%) strongly agreed with this statement.

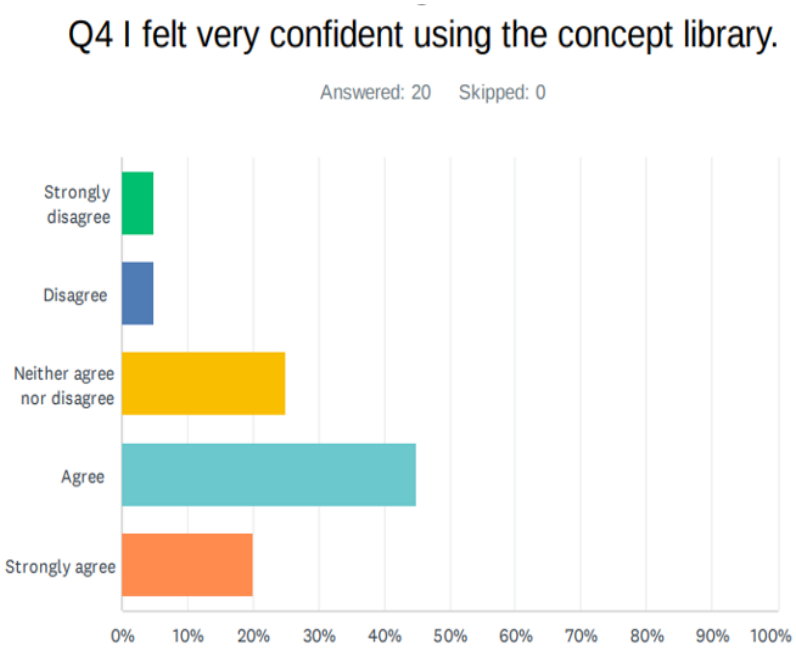
Figure 7.4 Responses of the participants to the third statement



7.5 RESPONSES OF PARTICIPANTS TO THE FOURTH STATEMENT: "I FELT VERY CONFIDENT USING THE CONCEPT LIBRARY"

The fourth statement's goal was to identify the percentage of participants who felt very confident about using the CALIBER research platform. The average score of this statement was 74%. The results in figure 7.5 showed that one participant (5.00%) strongly disagreed, one participant (5.00%) disagreed, five participants (25.00%) neither agreed nor disagreed, nine participants (45.00%) greed, and four participants (20.00%) strongly agreed with this statement.

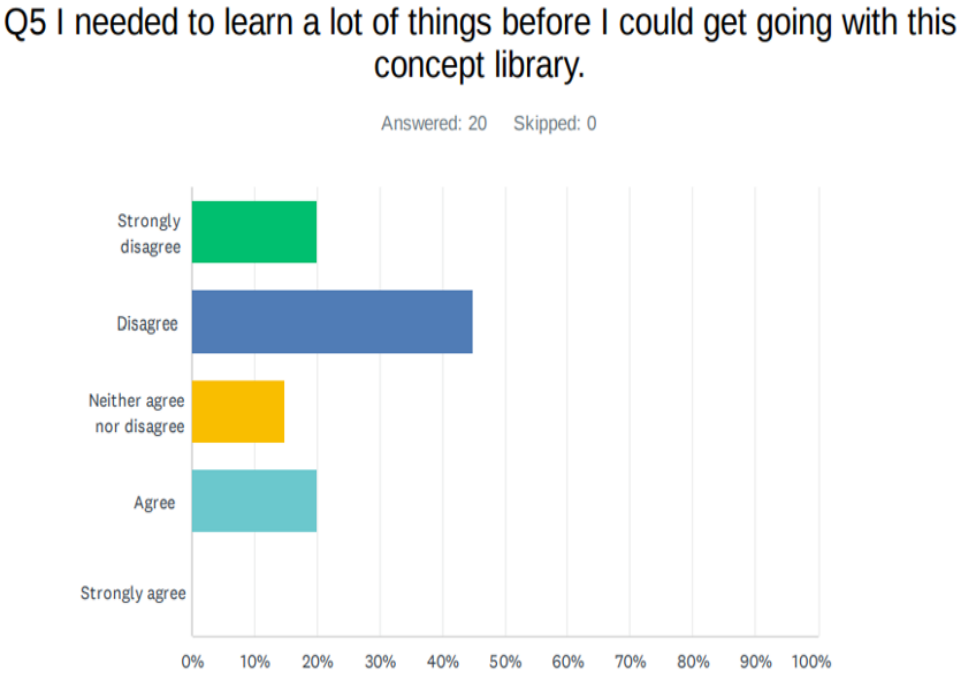
Figure 7.5 Percentages of the participants responses to the fourth statement



7.6 RESPONSES OF PARTICIPANTS TO THE FIFTH STATEMENT: "I NEEDED TO LEARN A LOT OF THINGS BEFORE I COULD GET GOING WITH THIS CONCEPT LIBRARY"

The purpose of the fifth statement was to measure the percentage of participants who needed to learn a lot of new information before they were able to get started with the CALIBER research platform. The average score of this statement was 73%. According to the results shown in figure 7.6 four participants (20.00%) strongly disagreed with this statement, nine participants (45.00%) disagreed, three participants (15.00%) were neither agreeing nor disagreeing, and four participants (20.00%) agreed with this statement.

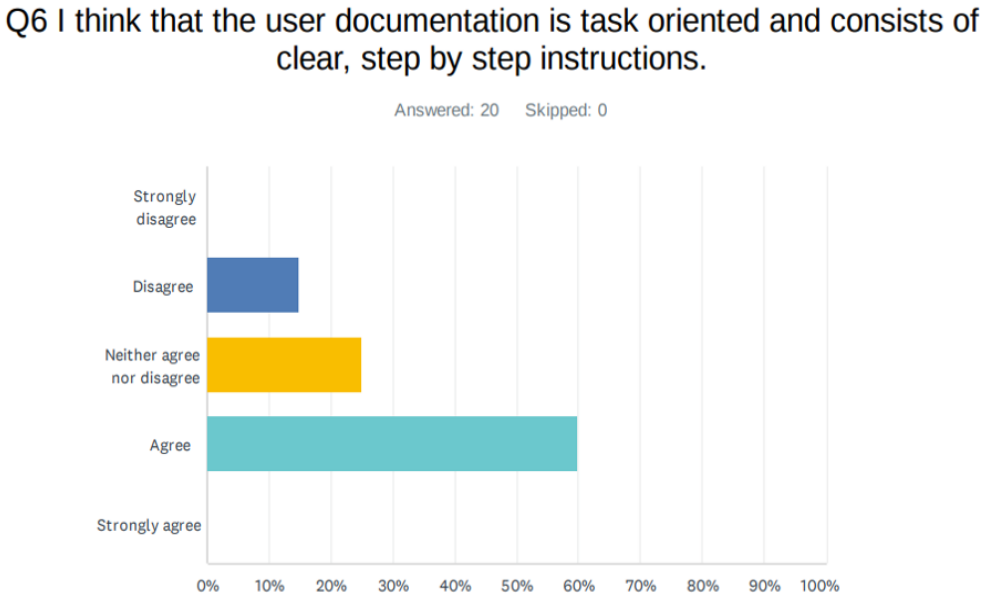
Figure 7.6 Percentages of the participants responses to the fifth statement



7.7 RESPONSES OF PARTICIPANTS TO THE SIXTH STATEMENT: "I THINK THAT THE USER DOCUMENTATION IS TASK ORIENTED AND CONSISTS OF CLEAR, STEP BY STEP INSTRUCTIONS"

The sixth statement was designed to determine the percentage of participants who believed the user documentation in the CALIBER research platform is task-oriented and includes clear, step-by-step instructions. The average score of this statement was 69%. According to the results shown in figure 7.7, three participants (15.00 %) disagreed, five (25.00 %) neither agreed nor disagreed, and twelve (60.00 %) agreed with this statement.

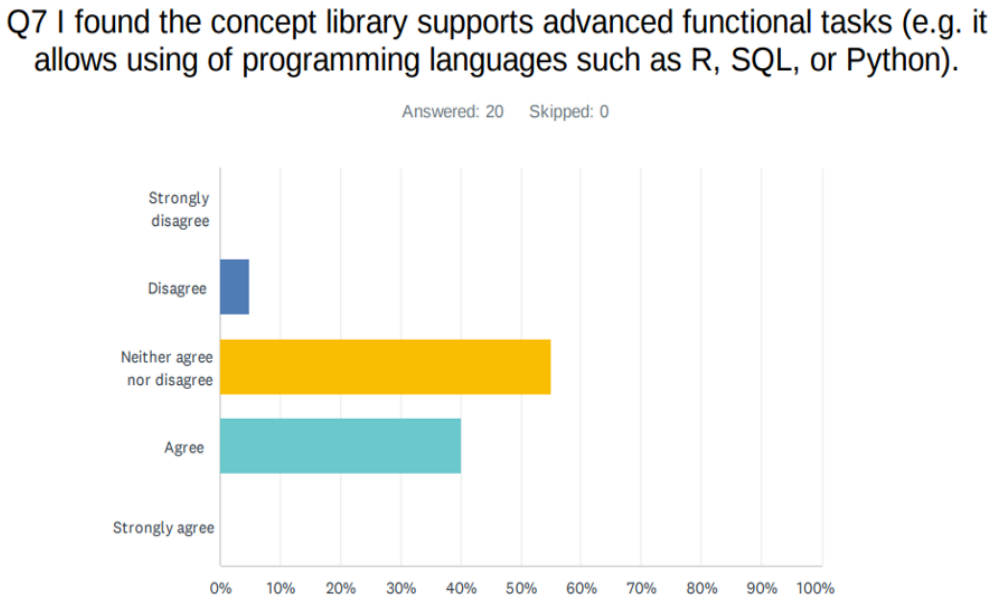
Figure 7.7 Percentages of the participants responses to the sixth statement



7.8 RESPONSES OF PARTICIPANTS TO THE SEVENTH STATEMENT: "I FOUND THE CONCEPT LIBRARY SUPPORTS ADVANCED FUNCTIONAL TASKS (E.G., IT ALLOWS USING OF PROGRAMMING LANGUAGES SUCH AS R, SQL, OR PYTHON)"

The purpose of the seventh statement was to determine the percentage of the participants who found the CALIBER research platform supported advanced functional tasks (e.g., it allows the use of programming languages such as R, SQL, or Python). The average score of this statement was 67%. The results in figure 7.8 showed that one participant (5.00%) disagreed, eleven participants (55.00%) neither agreed nor disagreed, and eight participants (40.00%) agreed with this statement.

Figure 7.8 Percentages of the participants responses to the seventh statement



7.9 RESPONSES OF PARTICIPANTS TO THE EIGHTH STATEMENT: "I FEEL IT IS ACCEPTABLE IF I AM REQUIRED TO REFERENCE THE CONCEPT LIBRARY WHEN PUBLISHING PAPERS"

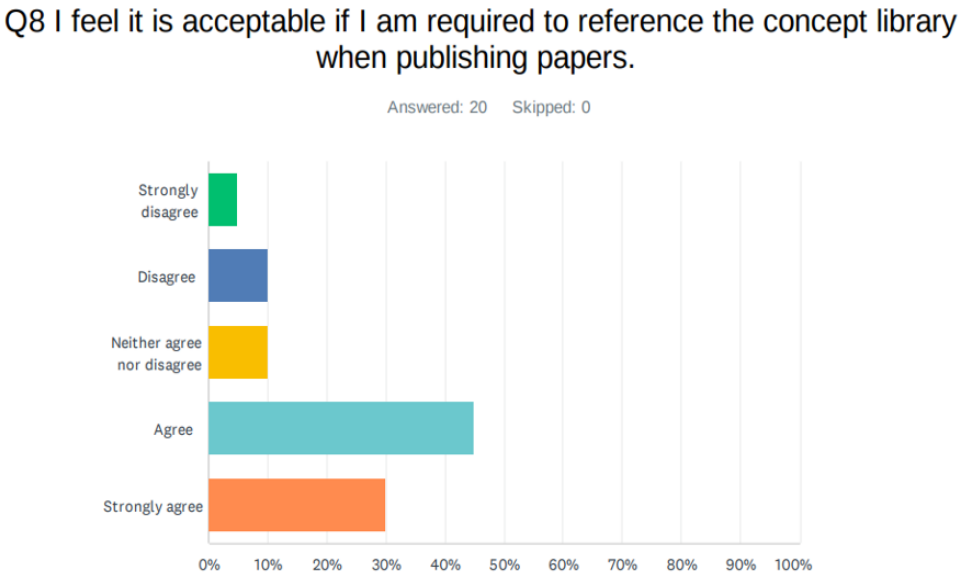
The eighth statement was designed to identify the percentage of participants who felt it was acceptable to reference the CALIBER research platform when publishing papers. The average score of this statement was 77%. Figure 7.9 showed that one participant (5.00%) strongly disagreed, two participants (10.00%) disagreed, and two participants (10.00%) neither agreed nor disagreed with this statement. The results also showed that Nine participants (45.00%) agreed with this statement for the following reasons:

- *"I don't mind referencing it if I use it"* (participant 2).
- *"I am more than happy to credit your excellent work!"* (participant 3).

However, six participants (30.00%) strongly agreed with this statement for the following reasons:

- *"One should acknowledge the source of codes and aggregations of codes"* (participant 4).
- *"At the very least the paper and modelist repository should be included as reference and noted within the methodology. It acknowledges the work of both the author and the concept library custodians in maintaining the code lists"* (participant 5).
- *"Makes our work more easily reproducible"* (participant 8).

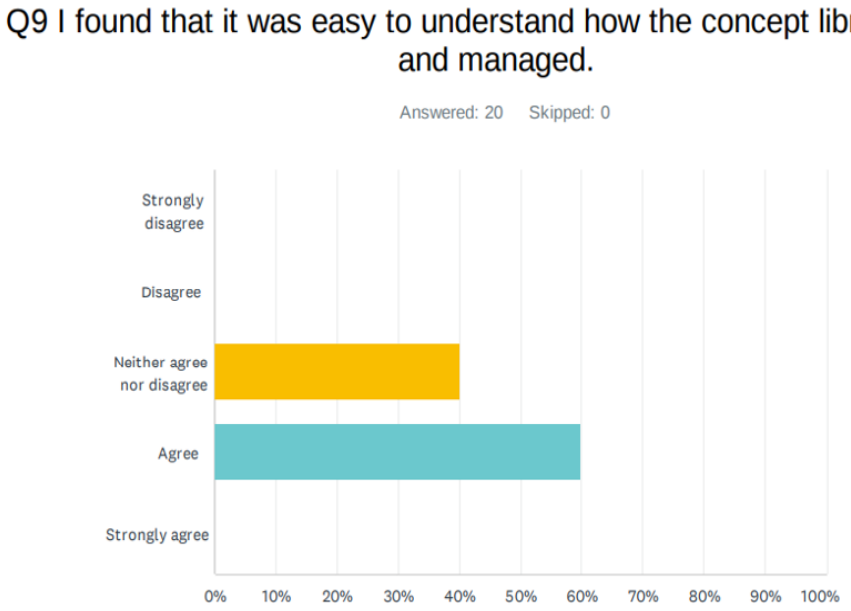
Figure 7.9 Percentages of the participants responses to the eighth statement



7.10 Responses of participants to the ninth statement: "I found that it was easy to understand how the concept library is run and managed"

The purpose of the ninth statement was to determine the percentage of the participants who found that it was easy to understand how the CALIBER research platform is run and managed. The average score of this statement was 72%. The results in figure 7.10 showed that eight participants (40.00%) neither agreed nor disagreed with this statement, and twelve participants (60.00%) agreed with this statement.

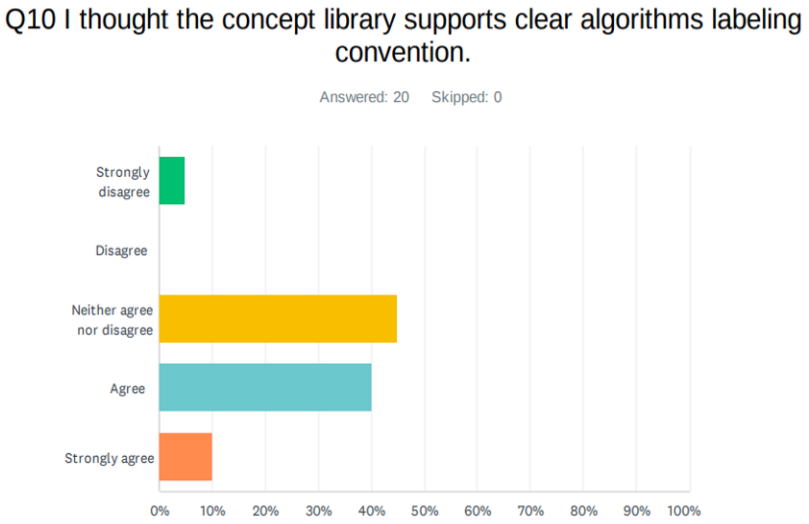
Figure 7.10 Percentages of the participants responses to the ninth statement



7.11 RESPONSES OF PARTICIPANTS TO THE TENTH STATEMENT: "I THOUGHT THE CONCEPT LIBRARY SUPPORTS CLEAR ALGORITHMS LABELLING CONVENTION"

Using the tenth statement, we were able to determine the percentage of participants who believed that the CALIBER research platform supports clear algorithm labelling conventions. The average score of this statement was 70%. According to the results shown in Figure 7.11, one participant (5.00%) strongly disagreed with this statement, nine participants (45.00%) neither agreed nor disagreed, eight participants (40.00%) agreed, and two participants (10.00%) strongly agreed with this statement.

Figure 7.11 Percentages of the participants responses to the tenth statement



7.12 Responses of participants to the eleventh statement: "Can you tell us more about why you give the answers you did? (e.g., what could be improved? and what did you like about the system?)"

The eleventh statement was created to encourage participants to freely express and share their opinions about the CALIBER research platform. The response rate to this statement was 100 percent. The majority of the participants' opinions were brief, and some of them were formed into sentences. The opinions of the participants were analysed using Braun and Clarke's six phases of thematic analysis [118]. Their opinions were read multiple times before being organised into themes and subthemes. Following this, the coding procedure began by giving equal importance to each opinion in order to ensure that the codes were comprehensive. Then, the emerging themes and subthemes from the data were assessed for relevance to the study, and related original themes were combined into a single one. Finally, the following labels were used to describe each theme and subtheme:

7.12.1 Suggestions for improving the quality of the CALIBER research platform:

1. Update the codes.
2. Enhance usability.
3. Support learnability.

There were several sub-themes that emerged from the first theme, which was about requesting improvements to the CALIBER research platform. One of the sub-themes was about improving the quality of the CALIBER research platform's content. Some participants suggested that the CALIBER lists need to be updated on an annual basis to include any additional code lists:

"I have 4 different versions of a psoriasis code lists when using CPRD primary care data – depending on its purpose in a given but why is only one listed in your repository? Probably because of a training and qualification issue. Or just a lack of experience? " (participant 11).

They also requested that the CALIBER research platform be updated to include the new SNOMED codes. Some participants suggested that the CALIBER research platform set a minimum standard for documenting the concepts. This would make the portal useful to most people. They also recommended integration with other system measurement purposes:

"I think integrating with BNF is important to accurately measure prescription exposures" (participant 20). Another subtheme suggested enhancing inquiries and facilitating the downloading of code lists. *"Downloadable CSV files would also be beneficial"* (participant 16).

A more specific request was for help and explanation for users, especially new users who have never used the portal before. This was so they could use it to its fullest potential.

7.12.2 Disadvantages that limit the usability of the CALIBER research platform:

1. The quality and validity of some content are uncertain.
2. The data is limited and context-specific.
3. The used classification systems are limited.
4. Lack of transparency in documenting the data.

The participants mentioned various drawbacks they experienced when using the CALIBER research platform. There were several sub-themes that emerged from this second theme. One sub theme was related to data accessibility and usability issues: *"It was not clear where to find and how to make use of the algorithms"* (participant 5). Some participants stated that when they found the codes, downloading them into Excel,

R, or Python was difficult*: "Getting the code lists into R is difficult (the package is a nightmare, sorry). I've used copy-paste in R to make this easier"*(participant 3).

Another sub-theme included concerns regarding the quality and validity of the CALIBER research platform's content. According to the participants, the code lists are not replicable, are limited to Read v2, and are context specific: "*It is not clear if it does include different versions of the diagnosis, both sensitive and specific ones"* (participant 10).

Some participants discovered that the CALIBER research platform provides a wide range of concepts in a single location. As a result, they may still refer to other published algorithms to learn more about the exclusion of certain codes or validation. Other participants stated that the CALIBER research platform is unsuitable for their research since it has limited code sets that do not cover a wide range of specialities (i.e., it mainly provides code lists of cardiovascular diseases). Some participants required a full and thorough understanding of many aspects (study type, design, purpose of the code lists, qualifications, in particular of the researchers, etc.) before considering using the code lists for their studies. However, some participants found out that common data modelling platforms such as atlas and aetion have made the conduct of EHRs studies too easy.

7.13 Responses of participants to the twelve statement: "If you are happy to participate in one-to-one interview, please provide us with your contact information"

The twelfth statement was intended to invite the participants to submit contact information if they were interested in participating in a one-to-one interview. 16 of the 20 participants skipped this statement, and just 4 gave contact information in order to participate in a one-to-one interview.

7.14 Interpreting the Concept library Usability Scale score

The Sauro and Lewis curved grading scale was used to interpret the Concept Library Usability Scale score [198] (See table 7.2). The average score of all participants (N=20) on the Concept Library Usability Scale was 72 (See figure 7.1), which corresponds to a grade of C+ and a 60–64 percentile range.

Table 7.2. The Sauro and Lewis curved grading scale

SUS score range	Grade	Percentile range
84.1–100	A+	96–100
80.8–84.0	A	90–95
78.9–80.7	A–	85–89
77.2–78.8	B+	80–84
74.1–77.1	B	70–79
72.6–74.0	B–	65–69
71.1–72.5	C+	60–64
65.0–71.0	C	41–59
62.7–64.9	C–	35–40
51.7–62.6	D	15–34
0.0–51.6	F	0–14

Source: (Sauro and Lewis, 2012, P204)

Chapter 8

General Discussion

In this chapter, I discuss all of the findings from my thesis work on the development of new concept libraries and the use of existing concept libraries, including the findings from the two review studies and from the three phases of this thesis: 1) the two qualitative studies, 2) the development of an email survey, and 3) the quantitative study with linked studies. Then I address the thesis's strengths, limitations, and original contributions to concept libraries. I also describe in this chapter the identified opportunities and obstacles from the users' perspectives in order to maximise the utility of concept libraries. Finally, I give recommendations for future research directions in order to improve reproducible research.

8 CHAPTER 8: DISCUSSION

8.1 SUMMARY OF FINDINGS

This thesis used an exploratory sequential mixed methods design in which the qualitative phase was dominant, implying that the qualitative phase was given more weight. This design was used so that survey data might help explain qualitative results for the goal of complementarity. The data were analysed, and the quantitative phase assisted in informing the qualitative phase.

Using the exploratory sequential mixed methods design in this thesis, allowed for a more in-depth exploration of the data, in those qualitative results from the interviews and focus group provided greater insights into why concept libraries are not be used effectively, as well as greater insights into quantitative survey results interpretation. In addition, it helped answer the following research question that cannot be answered just through quantitative or qualitative approaches: Do the perspectives of participants from the interviews, focus group, and survey instrument converge or differ? using quantitative or qualitative approaches alone would not provide appropriate answers to these questions.

As I mentioned in chapter one of my thesis, the main motivation for conducting this research in the United Kingdom was to pursue a PhD in a topic that could be useful for my home country, the State of Kuwait, which was in the process of integrating electronic data (the databases of primary health care centres with hospitals). As a result, I decided to learn more about data linking and the health informatics that surrounds it. I chose to investigate a concept library, which was relatively new in Wales. The analysis of the findings from all of the studies I conducted for my thesis, which included two qualitative studies (interviews and focus groups) and one quantitative study (a survey), revealed that the perspectives of participants from interviews, focus groups, and the survey instrument all agreed that concept libraries have several limitations and do not meet the various requirements of different users, and that they require lots of improvements in order to be useful and widely adopted. Therefore, I do not recommend that the State of Kuwait invest in the development of concept libraries.

8.2 ORIGINAL CONTRIBUTIONS

This thesis identified the requirements of a wide range of users from a variety of disciplines, including researchers, clinicians, machine learning experts, and research managers for a concept library. This thesis makes several contributions to existing knowledge on the subject. It provides an assessment of the usefulness of existing and newly proposed concept libraries from the user's perspective.

Two reviews of the literature are included in this thesis. The first review summarised various characteristics such as definitions, types, similarities, and differences of seven publicly accessible electronic health data concept libraries developed in different countries, including the United Kingdom, the United States of America, and Canada. This review may be of interest to researchers all over the world who want to learn more about the different characteristics of existing concept libraries. The second review summarised the clinical classification systems used to identify chronic diseases in children in routine data sources and other administrative datasets, including information on the coding systems upon which they are based (for example, ICD-10), the groupings used (for example, chapter headings within ICD-10), and, if possible, the specific codes used to identify a specific condition. This review could be useful to researchers who are interested in studies that combine a variety of data sources to classify a range of chronic conditions in children.

The qualitative studies (i.e., interviews and focus groups) identified and described in detail the needs of various users of a concept library. The findings of these studies could significantly improve the efficiency of existing concept libraries by informing their developers about the different needs and recommendations of various users. They would also help people who develop new concept libraries to improve access to and collaboration with routine EHR data by addressing the barriers and facilitators that have been identified in the qualitative studies. This would have an impact on other countries that want to access and share routine EHR data.

For the quantitative study, I developed the Concept Library Usability Scale, which is a specific measure that can be used to assess and compare the usability of concept libraries. It is a straightforward instrument that enables comprehensive assessments of the usability of concept libraries not just in the UK but globally. The Concept Library Usability Scale was made to be quick and easy to administer while still being reliable

enough to be used to measure the user experience of a concept library. I worked very hard to make it as short and simple as possible while also taking into account the ability of participants to complete the tasks of their choice, the effectiveness of those tasks' output, and the ease with which those tasks can be completed. This Concept Library Usability Scale, unlike the original SUS, includes comment boxes to obtain a thorough understanding of the participants' responses and to allow participants to explain why they agreed or disagreed with these statements.

8.3 INTERPRETATION OF FINDINGS IN THE LIGHT OF RELATED LITERATURE

8.3.1 Attitudes Toward the Development of A prototype Concept Library

The findings from the qualitative study presented in chapter 6, showed that although in principle, everyone felt that a digital portal containing a concept library would be very helpful, there were many requirements needed before its development. It needs to engage a wide variety of users if it is to be used (and current concept libraries are not widely used), which means it has to be very simple (point and click) for some, but it should have the software and usability to manipulate and design phenotyping algorithms for more advanced users. In addition, it needed to have a very high-quality search engine so that it is very easy to find information, and for it to expand, there needs to be a reason for users to upload their phenotyping algorithms, which needs to be very easy and quick.

This qualitative study indicated that although most of the interviewees expressed positive impressions about the idea of building a prototype concept library, approximately half of the participants expressed an interest in contributing to it. For the prototype concept library to work, researchers must engage with it and upload their codes there so that other people can use them. If researchers did not share their codes in the prototype concept library, this would usually mean an empty library. For better adoption of the prototype concept library, it is recommended that the developers consider the various facilitators for and barriers to participants sharing their work and reusing the work of others.

8.3.2 Facilitators and Barriers to Sharing and Reusing Research Methods

The findings of the focus group discussion (presented in chapter 6) demonstrate that facilitators of the participants' sharing of their research methods vary across four

categories: 1) personal drivers (e.g., obtaining appropriate credit, such as citations). This confirms the results of earlier studies that suggested researchers may be motivated to share their work if sharing leads to an increase in their citations [199] [200] [201], 2) benefits for their research team (e.g., sharing information to promote research within their team) [202] [203], 3) benefits for their organization (e.g., collaboration among researchers working within the same organization would advance their organization's research outcomes), and 4) benefits for the research community (e.g. expanding research base) [204]. With respect to this, Cragin et al have stated: *"As a research group gets larger and more formally connected to other research groups, it begins to function more like big science"* [205].

There were several barriers that could inhibit the participants from sharing their research methods, such as 1) the expected performance of the shared methods (e.g., some participants felt that building a general phenotyping algorithm to be used by others is very difficult). Similarly, Cragin et al have stated: *"Researchers could identify the data they thought was the most sharable, but not necessarily the most valuable for long-term preservation, particularly for reuse by other researchers"* [205], 2) lack of personal benefits such as recognition (e.g., some participants were worried about not being referenced by researchers who used their methods). In relation to this, Molloy reported that researchers can be discouraged from sharing their work by fear of not obtaining sufficient credit [206]. Therefore, a safeguard against uncredited use is necessary [207], 3) some participants mentioned that they were afraid that their methods would be used by other researchers as their own before publication. Similarly, the results of the study conducted by Huang et al. indicated that although most participants are interested in sharing papers related to biodiversity data, >60% of the participants were reluctant to share primary data before publication [208], and 4) some participants reported that lack adaptation of impact metrics may inhibit data sharing because researchers would not be able to measure the success of their methods if metrics are not available. This finding corresponds with other studies. For example, Costello stated *"citation services must include online data publications in their metrics"* [209], and Parr stated *"currently, authorship is considered paramount, particularly in journals with high impact factors based on overall journal citation rates. But a measure of individual researcher impact has recently been proposed"* [210].

Unless these obstacles are resolved, the sharing of data in concept libraries is unlikely to increase significantly.

Several facilitators encouraged participants to reuse research methods developed by others. They reported that reusing code lists created by other researchers would make their task much easier, save them a lot of time, and help to demonstrate that there is a justification for using such codes. These findings are consistent with those of the previous studies. For example, Anneke and Helen reported that researchers are using open research data to *"be aware of the state of the art and not recreate the wheel, as well as access to more data and generating fresh insights"* [211].

8.3.3 The Current System of Phenotyping

The results of the qualitative study (presented in chapter 6) indicate that more than half of the participants were not satisfied with their current system for phenotyping for several reasons, including the lack of accessibility of other researchers' work, such as code lists, which could affect research outcomes and the fact that reusing publicly available code lists consumes a lot of time and requires lots of work [204]; lack of confidence in web-based code lists if they are not cited by other researchers; lack of availability of a consistent approach for defining covariates such as smoking; and the selected Read code lists by the researchers are different from the selected Read code lists by the general practitioners. It seems that their current approach lacks confidence and is time-consuming and effort-intensive.

8.3.4 Implications and Potential Uses of Concept Libraries

The results of the qualitative study (presented in chapter 6) demonstrate that existing concept libraries are not widely used, and most participants who used some of the existing concept libraries expressed negative impressions about them (e.g., they do not provide training or user documentation, and they are difficult to use) [202] [203] [204]. Lack of knowledge of the existence of concept libraries and how to use them is generally described as an obstacle to data sharing [212]. As existing concept libraries are not used by all researchers, obstacles that inhibit researchers from using them need to be addressed when building new concept libraries.

8.3.5 The Concept Library Usability Scale

The purpose of developing the Concept Library Usability Scale, an e-mail survey instrument (See chapter 7), was to measure the usability of concept libraries by potential users. Therefore, a quantitative study was conducted, and the participants who work with the CPRD and SAIL Databank were asked to: 1) use one of the existing concept libraries in the UK, which is the CALIBER research platform, and 2) complete the Concept Library Usability Scale. The Concept Library Usability Scale was completed by only 20 of the 200 invited participants (i.e., 10%). Most researchers are interested in knowing what their acceptable survey response rate is. However, there is no agreement in the literature on what constitutes an acceptable response rate. For example, Lindemann posed the question, "*What is the average survey response rate?" The short answer? 33%."* [213]. Ramshaw mentioned, *"In our experience at Genroe, for well-crafted customer feedback surveys, the response rate is between 10% and 30%, depending on how engaged the audience is with the company"* [214], while Chung stated, *"A good survey response rate ranges between 5% and 30%"* [215].

Despite the fact that e-mail contact is the quickest and least expensive method of delivering this survey to participants, several factors may have influenced their low response rate, including: 1) the workload of professionals; 2) not including a deadline in the invitation emails; 3) different email checking habits (for example, some participants may open all of their emails, while others may only open emails from people they know; yet others may only open emails from their organization; and so on); 4) a lack of interest in the concept library idea; 5) different attitudes toward researchers (for example, some participants may be more likely to fill out a survey given by a PhD student, while others may be more likely to answer a survey given by a colleague in the field); and 6) the completing time (i.e., the extra 15 minutes that we asked the participants to spend using the CALIBER to search for algorithms they would like to use in the future before filling out the survey may inhibit some participants from completing the survey.)

Even those who responded were not very positive but had a lukewarm response. There is no clear support for the concept library idea. Although more than half of the participants agreed to use the CALIBER research platform for their studies, some

participants were not sure about its usefulness and thought that it had some limitations, such as not having many concepts. I identified patterns of variation in the perceptions of participants (N = 20) regarding the usability of the CALIBER research platform. For example, although more than half of the participants preferred to use the CALIBER research platform for their studies, several participants were not sure about its usefulness and mentioned some requirements related to the content of the portal and the people who created them:

1) Users required comprehensive information about the code lists, such as clarification of their purposes, to be available.
2) The code lists need to be updated to SNOMED CT as they are currently limited to Read codes, and conditions should be context specific.
3) Users wanted the portal to include more concepts and conditions.
4) Users desired standardised saved phenotypes to support reproducibility, reliability, and comparability.
5) The background of the people who made the code lists, such as their qualifications (e.g., having enough experience in the medical field like knowing medical terms and methods like different clinical classification systems), was important for users to know.

Accessibility to research data such code lists saved in concept libraries has significant potential for scientific advancement as it promotes the replication of research results and enables the use of old data in new contexts [212]. However, the results of this quantitative study indicated that some users felt the availability of the code lists in concept libraries alone was not enough to consider using them. Lack of transparency in documenting the entire process involved in creating the code lists and absence of knowledge about the impact of changes to code lists on the design of the study and interpretation of findings may inhibit the use of concept libraries.

In addition, the participants mentioned various drawbacks they experienced when using the CALIBER research platform related to data accessibility and usability issues, such as having difficulties locating and then using the algorithms, and when the code lists were located, downloading them into Excel, R, or Python was difficult. Also, the participants reported concerns regarding the quality and validity of the CALIBER

research platform's available code lists, such as that they are not replicable, limited to Read v2, and context-specific.

On the other hand, the results indicated that some users, who are probably new and non-experts, would use freely available data, including "fair" code lists, because of a lack of training, qualifications, and experience. They can easily link data and run studies but have very little understanding of what is actually happening with their data, including choices for creating code lists.

Some participants discovered that the CALIBER research platform provides a wide range of concepts in a single location. As a result, they may still refer to other published algorithms to learn more about the exclusion of certain codes or validation. Other participants stated that the CALIBER research platform is unsuitable for their research since it has limited code sets that do not cover a wide range of specialities (i.e., it mainly provides code lists of cardiovascular diseases). Some participants required a full and thorough understanding of many aspects (study type, design, purpose of the code lists, qualifications, in particular of the researchers, etc.) before considering using the code lists for their studies. Also, some participants said that they preferred to use common data modelling platforms, like atlas and aetion, to make their own algorithms to generate code lists from EHR routine data instead of using algorithms and code lists that other researcher have already made.

8.4 CHALLENGES

One of the most major difficulties I encountered while conducting my thesis was the unavailability of previous studies in the literature that explored the requirements of users for concept libraries. As a result, I decided to compare and contrast the findings of my thesis with those of previous studies that had addressed the sharing and reusing of research methods such as code lists. In addition, finding studies about a concept library for electronic health data phenotypes in the literature for the first review study was challenging because there are just a limited number of related studies (See chapter 4). The absence of a standard name or definition for concept libraries also resulted in a significant amount of time and effort being spent trying to locate them during the search process. In order to make inquiries as efficient as possible, a wider range of keywords had to be used.

Similarly, it was difficult to find relevant research papers for the second review study (See Chapter 5) because of the wide range of terminology, study types, and study designs in the literature. Conducting a thorough literature search on such a broad topic as chronic conditions in children was challenging, and some potentially eligible studies were excluded because they did not publish code lists.

Another challenge I experienced was difficulties in recruiting participants to complete the email survey, the Concept Library Usability Scale, for the quantitative study (See chapters 3 and 7). I invited (twice) 200 invited participants who work with the CPRD and SAIL Databank to complete the survey, and I kept the link to the survey open for about 4 months, but only 20 completed it.

8.5 Strengths and limitations

The first review study, presented in chapter 4, is the first review study, to our knowledge, aimed at identifying existing concept libraries, exploring their various characteristics, and examining the current practises in this evolving field. Finding studies about a concept library for electronic health data phenotypes in the literature was challenging as there were a limited number of related studies. Another challenge was the lack of a standard name or definition that describes this kind of library. Therefore, a range of keywords were needed to make queries as efficient as possible. This study examined only public concept libraries. It didn't look at non-publicly accessible concept libraries that are only accessible through the networks of the organisations and institutes that host them.

The second review study, presented in Chapter 5 of this thesis, aimed to identify the classification systems used for identifying all children with chronic conditions in routine data sources. This review was based on searching only one database (Medline), and this database was selected because it has the best coverage for health-related studies, which is where most studies in our area of interest would be indexed. The reference lists of all studies that had used a seemingly relevant classification system were also checked to ensure that any source studies were assessed against the eligibility criteria. Therefore, whilst it is possible that other relevant coding systems would have been identified if the searches had been repeated in other databases, I am reasonably confident that this search has identified most published papers that describe such systems. However, some judgement was required in selecting individual studies due

to the considerable differences in terminology, study types, and study designs found in the literature.

I feel that this review will be useful to researchers interested in identifying chronic conditions in children from routine data sources such as administrative databases. A summary of some of the available classification systems for chronic conditions in children, including links to code lists, would save researchers time and effort. For example, researchers may use the algorithms that use an ICD-based classification system to identify chronic conditions in data on hospitalisations or deaths in the UK, but only if the system matches or can be edited to fit their purposes. This review may also be useful for researchers interested in examining studies that classify a range of chronic conditions in children using a range of different combined data sources. However, additional work is needed to make sure that current systems that try to map ICD-10 codes to other systems for chronic conditions in children are validated.

This review may be beneficial for concept libraries such as clinicalcode.org and CALIBER [143] [18] as I tried to make the classification systems presented in this review repeatable by summarising the definitions, types of routine data sources, coding systems on which they are based (for example, ICD-10), and links to their specific codes if they are publicly available.

To our knowledge, the qualitative study provided in Chapter 6, is the first study aimed at identifying the needs of various users of a concept library. The findings of this study would have a significant impact on improving the efficiency of existing concept libraries by informing their developers about the different requirements, facilitators, barriers, and recommendations of the various users. In addition, this work will greatly inform the developers of new concept libraries to improve access to and collaboration with EHRs' routine data, which is part of an all-UK agenda, and the findings of this study will have implications for other countries working to access and share EHRs' routine data.

This qualitative study has some limitations that should be addressed in future studies. The first limitation is that we had a time limit on how long we could talk to the participants because each one-to-one interview was given 30 minutes. As a result, the number of questions we could ask and the amount of time we could spend on each

question were limited. The second limitation is that all the participants of the interviews and focus group discussion were recruited because they used the SAIL databank, a national eHealth data linkage infrastructure in Wales, so they mostly talked about the Swansea concept library in the SAIL databank. As the discussion focused on the SAIL databank, its generalization to other concept libraries was limited.

In the quantitative study presented in chapter 7, I developed the Concept Library Usability Scale, which is, to our knowledge, the first specific measure that can be used to assess and compare the usability of concept libraries. It is a straightforward instrument that enables comprehensive assessments of the usability of concept libraries globally. The Concept Library Usability Scale was made to be quick and easy to administer while still being reliable enough to be used to measure the user experience of a concept library. I worked very hard to make it as short and simple as possible while also taking into account the ability of participants to complete the tasks required, the effectiveness of those tasks' output, and the ease with which those tasks can be completed (See chapter 3).

This Concept Library Usability Scale, unlike the SUS, includes comment boxes. To obtain a thorough understanding of the participants' responses, I placed comment boxes below the following statements: 1, 2, and 8. This allowed participants to explain why they agreed or disagreed with these statements. To encourage participants to describe their overall responses, I have included an additional comment box following statement 11. Barnum said that the participants' comments, whether positive or negative, could help learn more about their experiences [137].

The low response rate of the participants (20 of 200 invited) who completed the survey is one of the limitations of my quantitative study presented in Chapter 7 because it could affect the interpretation of the results. Another limitation is that I only used My Concept Library Usability Scale to assess the usability of one concept library (the CALIBER data portal in the UK), which may limit the generalizability of the results. To overcome these limitations, I plan to use it in future research to see how well other concept libraries work.

8.6 FUTURE WORK

When I finish this thesis, I will conduct two studies. In the first study, I will add to the second review presented in Chapter 5 of this thesis, which aimed to identify classification systems used for identifying all children with chronic conditions in routine data sources, other studies that have tried to identify a limited sample of chronic conditions in children, such as most or the most common, or a specified set of chronic conditions (e.g., life-threatening conditions, children with disabilities), or a limited number of conditions. In addition, I will add other studies that have tried to modify other systems by adding in new codes or conditions. In the second study, I will use the Concept Library Usability Scale developed in this thesis to look at the usability of other concept libraries found in the literature, such as clinicalcode.org. I will then compare the new findings to the quantitative findings in this thesis.

Chapter 9

Conclusion

This chapter presents the general conclusion from all the studies conducted for this thesis, including the interviews, focus groups, and survey, and it provides recommendations to improve repeatable research using linked routine electronic data sources.

9 CHAPTER 9: CONCLUSION

Accessibility to research data such as code lists and algorithms has significant potential for scientific advancement as it promotes the replication of research results. The process of building code lists takes time and effort. Accordingly, access to existing code lists increases consistency and accuracy of definitions across studies, and collaboration between multiple research groups to develop validated code lists for studies with the same purposes is needed to avoid duplication of work. Also, transparency in sharing validated code lists and allowing other researchers to reuse them would be very useful for repeatable research for the public good.

Although it may seem beneficial for researchers to reuse methods developed by others, such as code lists, some researchers who created them prefer not to share them because they worked hard to create them and would rather publish them first to ensure their academic rights, such as being referenced. The major challenge is that some researchers would like to use the work of other researchers, but they do not want to contribute their work to concept libraries. Open sharing can be more difficult in the research community as researchers compete for grants, work promotions, and publication quotations. They think carefully about how, when, and where to share their work as they have spent a vast amount of time and effort to develop it. A solution to these issues would be to encourage researchers to contribute data to the prototype concept library in such a way that the shared data is understandable and reusable (e.g., ensuring uploading of adequate documentation) for the public good rather than for personal gains.

To enhance data sharing, the United Kingdom and other countries, including Canada, have built web-based data portals for phenotypes known as “concept libraries”, which allow data analysts, researchers, and clinicians to upload and download lists of clinical codes, update previous code lists, and share clinical code data across platforms. However, our review of the literature showed that the seven concept libraries identified were developed independently and appear to replicate similar concepts in different ways, including the ClinicalCodes.org [143], the Genotypes and Phenotypes Database (dbGaP) [146], the Phenotype knowledgebase (PheKB) [147], the Manitoba Centre for Health Policy (MCHP) Concept Dictionary and Glossary [148], the Clinical Disease Research using Linked Bespoke Studies and Electronic Health Records (CALIBER)

[19], the PhenoScanner V2 [149], and the Genome-Phenome Analysis Platform (GPAP) [150]. Furthermore, our review of the literature also revealed that there are a lot of different features in the different concept libraries, one concept library might do things that others do not (e.g., provides SNOMED or BNF code lists or provides definitions for demographic variables such as smoking or BMI algorithms that other concept libraries do not do). (See chapter 4)

Findings from qualitative and quantitative studies revealed that existing concept libraries are not widely used by researchers due to a number of limitations. (See chapters 6 and 7). To address these limitations, concept library developers may consider working together to: 1) raise awareness of their existence; 2) improve their different functions, such as enabling users to share, validate, and reuse concepts (e.g., code lists) to satisfy the various needs of their users; and 3) motivate users to contribute to those libraries. In addition, collaboration across data linkage centres is needed to develop common standards that govern and guide these emerging libraries. For example, they would agree on a relatively standard definition and name for concept libraries to enable users to locate them and then use them easily.

For a comprehensive adoption of concept libraries, their various functions, such as enabling users to share, validate, and reuse concepts (e.g., code lists), and their search features, should be assessed by developers, funders, users, and experts to ensure that they meet the needs of various users. Since there are two different types of concept libraries, 1) general libraries that hold phenotypes of multiple specialties and 2) specialised libraries that manage only certain specificities, such as rare diseases, users' preferences for the type of concept library need to be evaluated (e.g., through interviews, focus groups, and surveys) before developing new concept libraries.

The analysis of the findings from all of the studies that were conducted in this thesis, which included two qualitative studies (interviews and focus groups) and one quantitative study (a survey), revealed that although existing concept libraries have some limitations, they may have the potential to support repeatable research. To make concept libraries more useful, developers may consider investing in doing needs assessments for concept libraries in order to improve their functionalities and meet the expectations of various users. I recommend that the State of Kuwait begin by exploring existing concept libraries that provide code lists of interest to them, and then conduct

needs assessments for users before investing in the development of new concept libraries.

10 REFERENCES

1. Schleyer T, Song M, Gilbert GH, Rindal B, Fellows JL, Valeria V, et al. Electronic dental record use and clinical information management patterns among practitioner-investigators in The Dental Practice-Based Research Network. 2013; Available from: https://doi.org/10.14219/jada.archive.2013.0013

2. Wang SD. Opportunities and challenges of clinical research in the big-data era: From RCT to BCT. J Thorac Dis [Internet]. 2013;5(6):721–3. Available from: https://dx.doi.org/10.3978%2Fj.issn.2072-1439.2013.06.24

3. Pendergrass SA, Crawford DC. Using Electronic Health Records To Generate Phenotypes For Research. Curr Protoc Hum Genet [Internet]. 2019;100(1):1–20. Available from: https://doi.org/10.1002/cphg.80

4. Wei W, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes , clinical notes , and medications from electronic health records provides superior phenotyping performance. 2016;(January 2015):20–7. Available from: https://doi.org/10.1093/jamia/ocv130

5. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. 2013;117–21. Available from: https://doi.org/10.1136/amiajnl-2012-001145

6. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: A case study of atrial fibrillation. PLoS One. 2014;9(11).

7. Paraskevas Vezyridis ST. Open Access Research Evolution of primary care databases in UK: a scientometric analysis of research output. BMJ Open [Internet]. 2016; Available from: http://dx.doi.org/10.1136/bmjopen-2016-012785

8. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827–36.

9. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of the Health Improvement Network (THIN) database: Demographics, chronic disease prevalence and mortality rates. Inform Prim Care. 2011;19(4):251–5.

10. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: A new general practice database for research. Inform Prim Care. 2004;12(1):49–50.

11. Ford D V., 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 Serv Res [Internet]. 2009;9:1–12. Available from: https://doi.org/10.1186/1472-6963-9-157

12. Kousoulis AA, Rafi I, De Lusignan S. The CPRD and the RCGP: Building on research success by enhancing benefits for patients and practices. Br J Gen Pract. 2015;65(631):54–5.

13. Jones KH, Ford D V., Thompson S, Lyons RA. A profile of the SAIL databank on the UK secure research platform. Int J Popul Data Sci. 2019;4(2).

14. Lu M, Rupp LB, Trudeau S, Gordon SC. Validity of an automated algorithm using diagnosis and procedure codes to identify decompensated cirrhosis using electronic health records. 2017;369–76. Available from: https://doi.org/10.2147/clep.s136134

15. Cooksey R, Brophy S, Dennis M, Davies H, Atkinson M, Irvine E, et al. Severe flare as a predictor of poor outcome in ankylosing spondylitis: A cohort study using questionnaire and routine data linkage. Rheumatol (United Kingdom). 2015;54(9):1563–72.

16. Cooksey R, Husain MJ, Brophy S, Davies H, Rahman MA, Atkinson MD, et al. The cost of ankylosing spondylitis in the UK using linked routine and patient-reported survey data. PLoS One. 2015;10(7):1–17.

17. UCL Institute of Health Informatics. CALIBER [Internet]. [cited 2021 Apr 21]. Available from: https://www.ucl.ac.uk/health-informatics/caliber

18. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al.

Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012;41(6):1625–38.

19. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019;26(12):1545–59.

20. 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. Eur J Epidemiol [Internet]. 2019;34(1):91–9. Available from: https://doi.org/10.1007/s10654-018-0442-4

21. Herrett E, Smeeth L, Walker L, Weston C. The Myocardial Ischaemia National Audit Project (MINAP). Heart [Internet]. 2010;96(16):1264–7. Available from: http://dx.doi.org/10.1136/hrt.2009.192328

22. NHS Digital. Hospital Episode Statistics (HES) [Internet]. [cited 2021 Apr 22]. Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

23. Office for National Statistics. About us [Internet]. [cited 2021 Apr 22]. Available from: https://www.ons.gov.uk/

24. NHS Digital [Internet]. [cited 2021 Apr 25]. Available from: https://digital.nhs.uk/

25. George J, Herrett E, Denaxas S, Nicholas O, Shah A, Rapsomaniki E, et al. Differential Effects of Smoking on Specific Cardiovascular Presentations in Men and Women: Prospective Cohort Study in 900,000 Patients Using CALIBER Linked Electronic Health Records. Am Hear Assoc Inc. 2012;

26. Chung SC, Gedeborg R, Nicholas O, James S, Jeppsson A, Deanfield J, et al. Comparative Effectiveness of Acute Myocardial Infarction Care Delivered In Sweden and the United Kingdom Using National Outcome Registries. Am

Hear Assoc Inc. 2012;

27. Rapsomaniki E, Shah A, Perel P, Denaxas S, George J, Nicholas O, et al. Prognostic models for stable coronary artery disease based on electronic health record cohort of 102 023 patients. Eur Heart J. 2014;35(13):844–52.

28. CPRD UK data driving real-world evidence. (n.d para 1.). Clinical Practice Research Datalink. Retrieved January 13, 2021 from https://cprd.com/home [Internet]. [cited 2021 Jan 13]. Available from: https://cprd.com/home

29. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, Van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: Cohort study. BMJ [Internet]. 2013;346(7909):1–12. Available from: http://dx.doi.org/doi:10.1136/bmj.f2350

30. McGowan VF. Introduction to Medical Coding, Introductory Module. 2019; Available from: https://icd.who.int/

31. Terminology and Classifications Delivery Service. National Clinical Coding Standards OPCS-4 (2021). 2021;

32. Pohontsch NJ, Zimmermann T, Jonas C, Lehmann M, Löwe B, Scherer M. Coding of medically unexplained symptoms and somatoform disorders by general practitioners - An exploratory focus group study. BMC Fam Pract. 2018;19(1):1–11.

33. Alakrawi ZM. Clinical Terminology and Clinical Classification Systems: A Critique Using AHIMA's Data Quality Management Model. Prespectives Heal Inf Manag [Internet]. 2016; Available from: http://perspectives.ahima.org/clinical-terminology-and-clinical-classification-systems-a-critique/

34. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. J Biomed Inform [Internet]. 2017;70:1–13. Available from: http://dx.doi.org/10.1016/j.jbi.2017.04.010

35. Imel M, Campbell J. Mapping from a Clinical Terminology to a Classification. AHIMA's 75th Anniv Natl Conv … [Internet]. 2003; Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Mapping+from+a+Clinical+Terminology+to+a+Classification#0

36. NHS Digital. User Guide SCCI0034: SNOMED CT. Computer (Long Beach Calif) [Internet]. 2020;(September):169–232. Available from: www.impact-test.co.uk

37. NHS Digital. SNOMED CT [Internet]. [cited 2021 Feb 7]. Available from: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct

38. Scottish Clinical Information Management in Practice. Read Terms for General Practice. Available from: https://www.scimp.scot.nhs.uk/better-information/clinical-coding/scimp-guide-to-read-codes

39. NHS Digital. Read Codes [Internet]. [cited 2021 Feb 27]. Available from: https://digital.nhs.uk/services/terminology-and-classifications/read-codes

40. NHS Digital. Guidance for Primary Care: Transitioning from Read to SNOMED CT. 2016;(December):23. Available from: https://www.networks.nhs.uk/nhs-networks/snomed-ct/snomed-ct-learning/guidance-for-primary-care-transitioning-from-read-to-snomed-ct

41. Brogan T. Medical Terminologies and Classification Systems. Heal Inf Technol basics a concise Guid to Princ Pract [Internet]. 2009;103–16. Available from: http://www.jblearning.com/samples/0763746878/46878_fmxx_pass1rev.pdf

42. World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD) [Internet]. [cited 2021 Feb 10]. Available from: https://www.who.int/classifications/classification-of-diseases

43. NHS Digital. Terminology and Classifications [Internet]. [cited 2021 Mar 3]. Available from: https://digital.nhs.uk/services/terminology-and-classifications

44. Health T. National Interim Clinical Imaging Procedure (NICIP) Code Set -

Frequently Asked Questions. 2017;(April):1–6.

45. Digital NHS. National Interim Clinical Imaging Procedure (NICIP) Code Set – Implementation Guidance. 2019;(June):1–20.

46. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annu Rev Biomed Data Sci. 2018;1(1):53–68.

47. Alzoubi H, Alzubi R, Ramzan N, West D, Al-hadhrami T. A Review of Automatic Phenotyping Approaches using Electronic Health Records. :1–23.

48. Mo H, Thompson WK, Rasmussen L V., Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Informatics Assoc. 2015;22(6):1220–30.

49. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Informatics Assoc. 2014;21(2):221–30.

50. Michalik DE, Taylor BW, Panepinto JA. Identification and Validation of a Sickle Cell Disease Cohort Within Electronic Health Records. Acad Pediatr [Internet]. 2017;17(3):283–7. Available from: http://dx.doi.org/10.1016/j.acap.2016.12.005

51. Nguyen AN, Lawley MJ, Hansen DP, Bowman R V., Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Informatics Assoc. 2010;17(4):440–5.

52. Restrepo NA, Farber-Eger E, Crawford DC. Searching in the Dark: Phenotyping Diabetic Retinopathy in a De-Identified Electronic Medical Record Sample of African Americans. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci [Internet]. 2016;2016:221–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27570675%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5001772

53. Aref-Eshghi E, Oake J, Godwin M, Aubrey-Bassler K, Duke P, Mahdavian M, et al. Identification of Dyslipidemic Patients Attending Primary Care Clinics Using Electronic Medical Record (EMR) Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Database. J Med Syst. 2017;41(3).

54. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Informatics Assoc. 2012;19(2):212–8.

55. Safarova MS, Liu H, Kullo IJ. Rapid identification of familial hypercholesterolemia from electronic health records: The SEARCH study. J Clin Lipidol [Internet]. 2016;10(5):1230–9. Available from: http://dx.doi.org/10.1016/j.jacl.2016.08.001

56. Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. Clin Pharmacol Ther. 2020;107(4):871–85.

57. Bisong E. Logistic Regression. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. 2019.

58. Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018;19(Suppl 17).

59. Navada A, Ansari AN, Patil S, Sonkamble BA. Overview of use of decision tree algorithms in machine learning. Proc - 2011 IEEE Control Syst Grad Res Colloquium, ICSGRC 2011. 2011;37–42.

60. Zhou SM, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis. PLoS One. 2016;11(5):1–14.

61. Wu ST, Sohn S, Ravikumar KE, Wagholikar K, Jonnalagadda SR, Liu H, et

al. Automated chart review for asthma cohort identification using natural language processing: An exploratory study. Ann Allergy, Asthma Immunol. 2013;111(5):364–9.

62. Weissler EH, Lippmann SJ, Smerek MM, Ward RA, Kansal A, Brock A, et al. Model-Based Algorithms for Detecting Peripheral Artery Disease Using Administrative Data From an Electronic Health Record Data System: Algorithm Development Study. JMIR Med Informatics. 2020;8(8):e18542.

63. Jorge A, Castro VM, Barnado A, Gainer V, Hong C, Cai T, et al. Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms. Semin Arthritis Rheum [Internet]. 2019;49(1):84–90. Available from: https://doi.org/10.1016/j.semarthrit.2019.01.002

64. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: High-throughput candidate phenotype generation via tensor factorization. J Biomed Inform [Internet]. 2014;52:199–211. Available from: http://dx.doi.org/10.1016/j.jbi.2014.07.001

65. Kim Y, El-Kareh R, Sun J, Yu H, Jiang X. Discriminative and Distinct Phenotyping by Constrained Tensor Factorization. Sci Rep [Internet]. 2017;7(1):1–12. Available from: http://dx.doi.org/10.1038/s41598-017-01139-y

66. Perros I, Papalexakis EE, Vuduc R, Searles E, Sun J. Temporal phenotyping of medically complex children via PARAFAC2 tensor factorization. J Biomed Inform [Internet]. 2019;93(September 2018):103125. Available from: https://doi.org/10.1016/j.jbi.2019.103125

67. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak. 2019;19(1):1–16.

68. Panahiazar M, Taslimitehrani V, Pereira NL, Pathak J. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. Stud Health Technol Inform. 2015;210:369–73.

69. Kagawa R, Kawazoe Y, Ida Y, Shinohara E, Tanaka K, Imai T, et al. Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach. J Diabetes Sci Technol. 2017;11(4):791–9.

70. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. Arthritis Res Ther. 2019;21(1):1–9.

71. Richesson R, Smerek M. Electronic Health Records-Based Phenotyping. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. [Internet]. [cited 2020 Dec 3]. Available from: http://sites.duke.edu/rethinkingclinicaltrials/informed-consent-in-pragmatic-clinical-trials/

72. Deleger L, Lingren T, Ni Y, Kaiser M, Stoutenborough L, Marsolo K, et al. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. J Biomed Inform [Internet]. 2014;50:173–83. Available from: http://dx.doi.org/10.1016/j.jbi.2014.01.014

73. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. J Am Med Informatics Assoc. 2013;20(E2).

74. Baldiviez LM, Keim NL, Laugero KD, Hwang DH, Huang L, Woodhouse LR, et al. Design and implementation of a cross-sectional nutritional phenotyping study in healthy US adults. BMC Nutr. 2017;3(1):1–13.

75. Suojalehto H, Suuronen K, Cullinan P, Lindström I, Sastre J, Walusiak-Skorupa J, et al. Phenotyping Occupational Asthma Caused by Acrylates in a Multicenter Cohort Study. J Allergy Clin Immunol Pract. 2020;8(3):971-979.e1.

76. LePendu P, Iyer S V., Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. J Biomed Semantics. 2012;3(1):1–12.

77. Franks PW, Pomares-Millan H. Next-generation epidemiology: the role of high-resolution molecular phenotyping in diabetes research. Diabetologia. 2020;2521–32.

78. Norcliffe-Kaufmann L, Slaugenhaupt SA, Kaufmann H. Familial dysautonomia: History, genotype, phenotype and translational research. Prog Neurobiol [Internet]. 2017;152:131–48. Available from: http://dx.doi.org/10.1016/j.pneurobio.2016.06.003

79. Wolford BN, Willer CJ, Surakka I. Electronic health records: The next wave of complex disease genetics. Hum Mol Genet. 2018;27(R1):R14–21.

80. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: A perspective from the NIH health care systems collaboratory. J Am Med Informatics Assoc. 2013;20(E2).

81. Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG. Archetype-based data warehouse environment to enable the reuse of electronic health record data. Int J Med Inform [Internet]. 2015;84(9):702–14. Available from: http://dx.doi.org/10.1016/j.ijmedinf.2015.05.016

82. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: What information is hidden in free text? BMC Med Res Methodol [Internet]. 2013;13(1):1. Available from: BMC Medical Research Methodology

83. Barnado A, Casey C, Carroll RJ, Wheless L, Denny JC, Crofford LJ. Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus. Arthritis Care Res. 2017;69(5):687–93.

84. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. Yearb Med Inform. 2014;9:215–23.

85. Brandt PS, Kiefer RC, Pacheco JA, Adekkanattu P, Sholle ET, Ahmad FS, et

al. Toward cross-platform electronic health record-driven phenotyping using Clinical Quality Language. Learn Heal Syst. 2020;4(4):1–9.

86. Bozkurt S, Paul R, Coquet J, Sun R, Banerjee I, Brooks JD, et al. Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. Learn Heal Syst. 2020;4(4):1–9.

87. Akbarov A, Kontopantelis E, Sperrin M, Stocks SJ, Williams R, Rodgers S, et al. Primary Care Medication Safety Surveillance with Integrated Primary and Secondary Care Electronic Health Records: A Cross-Sectional Study. Drug Saf. 2015;38(7):671–82.

88. Manuel DG, Rosella LC ST. Importance of accurately identifying disease in studies using electronic health records. BMJ [Internet]. 2010;341(7770):443. Available from: https://doi.org/10.1136/bmj.c4226

89. Nicholson A, Tate AR, Koeling R CJ. What does validation of cases in electronic record databases mean? The potential contribution of free text. Wiley Online Libr [Internet]. 2011;(Ci):321–4. Available from: https://doi.org/10.1002/pds.2086

90. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes : An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. 2014;9(6):6–11. Available from: https://doi.org/10.1371/journal.pone.0099825

91. Bhattarai N, Charlton J, Rudisill C, Gulliford MC. Coding, recording and incidence of different forms of coronary heart disease in primary care. PLoS One [Internet]. 2012;7(1). Available from: https://doi.org/10.1371/journal.pone.0029776

92. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM. Selection of Medical Diagnostic Codes for Analysis of Electronic Patient Records. Application to Stroke in a Primary Care Database. PLoS One [Internet]. 2009;4(9):e7168. Available from: https://doi.org/10.1371/journal.pone.0007168

93. Sargeant JM, O'connor AM, Dohoo IR, Erb HN, Cevallos M, Egger M, et al. Methods and processes of developing the strengthening the reporting of observational studies in epidemiology-veterinary (STROBE-Vet) statement. J Food Prot [Internet]. 2016;79(12):2211–9. Available from: https://doi.org/10.1111/jvim.14574

94. Harron K, Benchimol E, Langan S. Using the RECORD guidelines to improve transparent reporting of studies based on routinely collected data. Int J Popul Data Sci [Internet]. 2018;3(1):10–3. Available from: https://dx.doi.org/10.23889%2Fijpds.v3i1.419

95. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Peteresen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med [Internet]. 2015;12(10):1–18. Available from: https://doi.org/10.1371/journal.pmed.1001885

96. Ostapyk T. Manitoba Centre for Health Policy Data Repository. In: Michalos AC (eds) Encyclopedia of Quality of Life and Well-Being Research [Internet]. Springer, Dordrecht; 2014. Available from: https://doi.org/10.1007/978-94-007-0753-5_3483

97. Doyle L, Brady AM, Byrne G. An overview of mixed methods research – revisited. J Res Nurs. 2016;21(8):623–35.

98. Sockolow P, Dowding D, Randell R, Favela J. Using mixed methods in health information technology evaluation. Stud Health Technol Inform. 2016;225:83–7.

99. Creswell JW, Clark VLP. Designing and conducting mixed methods research. Thousand Oaks, CA SAGE. 2017;Third Edit.

100. Johnson RB, Onwuegbuzie AJ, Turner LA. Toward a Definition of Mixed Methods Research. J Mix Methods Res [Internet]. 2007; Available from: https://doi.org/10.1177/1558689806298224

101. Shorten A, Smith J. Mixed methods research: Expanding the evidence base. Evid Based Nurs. 2017;20(3):74–5.

102. Halcomb EJ, Hickman L. Mixed methods research. Nurs Stand Promot Excell Nurs care [Internet]. 2015;29(32):41–7. Available from: https://ro.uow.edu.au/smhpapers/2656

103. Greene JC, Caracelli VJ, Graham WF. Toward a Conceptual Framework for Mixed-Method Evaluation Designs. Educ Eval Policy Anal. 1989;11(3):255–74.

104. Bryman A. Integrating quantitative and qualitative research: How is it done? Qual Res. 2006;6(1):97–113.

105. Creswell JWL, Clark VP. Choosing a mixed methods design. Des Conduct Mix methods Res [Internet]. 2010;53–106. Available from: http://www.sagepub.com/sites/default/files/upm-binaries/35066_Chapter3.pdf

106. Clark VLP, Ivankova N V. How do Personal Contexts Shape Mixed Methods?: Considering Philosophical, Theoretical, and Experiential Foundations for Mixed Methods Research. Mix Methods Res A Guid to F. 2018;191–216.

107. Braun V, Clarke V. What can "thematic analysis" offer health and wellbeing researchers? Int J Qual Stud Health Well-being. 2014;9:9–10.

108. Oates J, Carpenter D, Fisher M, Goodson S, Beth H, Kwiatowski R, et al. BPS Code of Human Research Ethics [Internet]. 2nd ed. British Psychological Society; 2021. 42 p. Available from: https://www.bps.org.uk/sites/www.bps.org.uk/files/Policy/Policy - Files/BPS Code of Human Research Ethics.pdf

109. Halej J. Ethics in primary research (focus groups, interviews and surveys). ECU Res Data Brief [Internet]. 2017;1–13. Available from: https://warwick.ac.uk/fac/cross_fac/ias/activities/accolade/resources/ecu_research_ethics.pdf

110. The Department of Human Resources and Change Information Rights and Information Security (IRIS) Service. Data Protection Act 1998: Personal information about constituents and others Advice for Members and their staff.

1998;

111. Ryan F, Coughlan M, Cronin P. Interviewing in qualitative research: The one-to-one interview. Int J Ther Rehabil. 2009;16(6):309–14.

112. Mathers N, Nick Fox AH. Trent Focus for Research and Development in Primary Health Care. Inst Gen Pract North Gen Hosp Sheff. 1998;26(11):1–34.

113. Lambert SD, Loiselle CG. Combining individual interviews and focus groups to enhance data richness. J Adv Nurs. 2008;62(2):228–37.

114. Stofer KA. Preparing for One-on-One Qualitative Interviews: Designing and Conducting the Interview. Edis. 2019;2019(4):4.

115. Boyce C, Neale P. Conducting In-Depth Interviews: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input. Pathfind Int. 2006;2(May):1–16.

116. McQuarrie EF, Krueger RA. Focus Groups: A Practical Guide for Applied Research. J Mark Res [Internet]. 3rd ed. 1989;26(3):371. Available from: https://doi.org/10.2307/3172912

117. Onwuegbuzie AJ, Dickinson WB, Leech NL, Zoran AG. A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research. Int J Qual Methods [Internet]. 2009;8(3):1–21. Available from: https://doi.org/10.1177%2F160940690900800301

118. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol [Internet]. 2006;3(2):77–101. Available from: https://doi.org/10.1191/1478088706qp063oa

119. O.Nyumba T, Wilson K, Derrick CJ, Mukherjee N. The use of focus group discussion methodology: Insights from two decades of application in conservation. Methods Ecol Evol. 2018;9(1):20–32.

120. Hennink, M. Monique. Focus group discussions [Internet]. Oxford University Press. 2014. Available from:

https://doi.org/10.1093/acprof:osobl/9780199856169.001.0001

121. Smith F. Focus groups. Pharm Pract Second Ed. 2017;407–16.

122. EU Joint Action on Chronic Diseases and Healthy Ageing Across the Life Cycle (JA-CHRODIS). JA-CHRODIS Work Package 7 Diabetes : a case study on strengthening health care for people with chronic diseases SWOT ANALYSIS Overview of national or sub national policies and programs on prevention and management of diabetes. 2016;1–79.

123. Sarsby A. A Useful Guide to SWOT Analysis [Internet]. 2012. Available from: http://www.cii.co.uk/media/6158020/a-useful-guide-to-swot-analysis.pdf

124. Gürel E. SWOT ANALYSIS: A THEORETICAL REVIEW. J Int Soc Res. 2017;10(51).

125. Namugenyi C, Nimmagadda SL, Reiners T. Design of a SWOT analysis model and its evaluation in diverse digital business ecosystem contexts. Procedia Comput Sci [Internet]. 2019;159:1145–54. Available from: https://doi.org/10.1016/j.procs.2019.09.283

126. Sammut-Bonnici T, Galea D. SWOT Analysis. Wiley Encycl Manag. 2015;(January).

127. Thayer D, Bown D, Leake T, Jones J-L, Noyce R, Brooks C, et al. Code List Library: A Solution to Improve Research Repeatability, Transparency, and Efficiency by Curating Lists of Clinical Codes. Int J Popul Data Sci [Internet]. 2018;0(September):23889. Available from: https://doi.org/10.23889/ijpds.v3i4.891

128. Braun V, Clarke V. Thematic analysis. APA Handb Res methods Psychol Vol 2 Res Des Quant Qual Neuropsychol Biol. 2012;2:57–71.

129. Guest G, MacQueen K, Namey E. Introduction to Applied Thematic Analysis. Appl Themat Anal. 2014;3–20.

130. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No.

131. Med Teach [Internet]. 2020;42(8):846–54. Available from: https://doi.org/10.1080/0142159X.2020.1755030

131. Braun and Clarke. Successful Qualitative Research: A Practical Guide for Beginners - Virginia Braun, Victoria Clarke - Google Books. 2013;382.

132. Workbook. Introduction to Using nVivo. 2017;(January):1–15.

133. Toepoel V. Introduction to using online surveys. In: Doing Surveys Online. 55 City Road: SAGE Publications Ltd; 2017. p. 1–18.

134. Geldsetzer P. Use of rapid online surveys to assess people's perceptions during infectious disease outbreaks: A Cross-sectional Survey on COVID-19. J Med Internet Res. 2020;22(4):1–13.

135. Epstein L. Get better data and simplify analysis: Qualitative vs. quantitative questions [Internet]. [cited 2021 Aug 16]. Available from: https://www.surveymonkey.com/curiosity/qualitative-vs-quantitative/

136. Sue V, Ritter L. Planning the Online Survey. In: Conducting Online Surveys. Edition, 2. Thousand Oaks: SAGE Publications, Inc; 2015. p. 14–32.

137. Barnum CM. Usability Testing Essentials: Ready, Set… Test! 1st ed. Burlington: MA, Morgan Kaufmann Publishers; 2011.

138. Sauro J, Lewis JR. Chapter 8 - Standardized usability questionnaires in Quantifying the User Experience (Second Edition) [Internet]. 2016. 185–248 p. Available from: http://www.sciencedirect.com/science/article/pii/B9780128023082000084

139. Brooke J. SUS: A "Quick and Dirty" Usability Scale. Usability Eval Ind. 2020;207–12.

140. Lewis JR. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. Int J Hum Comput Interact. 1995;7(1):57–78.

141. McLeod S. Likert Scale Likert Scale Examples How can you analyze data

from a Likert Scale ? Simply Psychol [Internet]. 2008;1–2. Available from: https://www.simplypsychology.org/likert-scale.html

142. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: Results and lessons learned from the eMERGE network. J Am Med Informatics Assoc. 2013;20(E1):147–54.

143. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. PLoS One [Internet]. 2014;9(6):6–11. Available from: https://doi.org/10.1371/journal.pone.0099825

144. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: A case study of atrial fibrillation. PLoS One [Internet]. 2014;9(11). Available from: https://doi.org/10.1371/journal.pone.0110900

145. Arksey H, O'Malley L. Scoping studies: Towards a methodological framework. Int J Soc Res Methodol Theory Pract. 2005;8(1):19–32.

146. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's database of genotypes and phenotypes: DbGaP. Nucleic Acids Res. 2014;42(D1):975–9.

147. Kirby JC, Speltz P, Rasmussen L V., Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Informatics Assoc. 2016;23(6):1046–52.

148. University of Manitoba. Concept dictionary and glossary for population based research (n.d. para. 1, 2) [Internet]. [cited 2020 Jan 29]. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/concept_dictionary.html

149. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al.

PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019;35(22):4851–3.

150. RD-Connect. Genome-Phenome Analysis Platform GPAP (n.d. para 1) [Internet]. [cited 2020 Jan 31]. Available from: https://platform.rd-connect.eu/

151. Shah A. CALIBERcodelists user guide. 2014;1–23. Available from: https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/CALIBERcodelists/inst/doc/user guide.pdf?root=caliberanalysis

152. Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014;29(SUPPL. 3):780–7.

153. RD-Connect. Genome-Phenome Analysis Platform (n.d. para 1, 4) [Internet]. [cited 2020 Jan 31]. Available from: https://rd-connect.eu/what-we-do/omics/gpap/

154. Smith, M1, Turner, K1, Bond, R1, Kawakami, T1, and Roos L. The Concept Dictionary and Glossary at MCHP: Tools and Techniques to Sup- port a Population Research Data Repository. 2019;0(December):1–4.

155. RD-Connect. Bioinformatic Tools [Internet]. Available from: https://rd-connect.eu/what-we-do/bioinformatic-tools/

156. dbGaP Genotypes and Phenotypes. dbGaP Overview (n.d. para.1) [Internet]. [cited 2020 Jan 30]. Available from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html

157. Clinicalcodes.org. Clinical codes [Internet]. [cited 2020 Jan 28]. Available from: https://clinicalcodes.rss.mhs.man.ac.uk/

158. PhenoScanner V2. Licence: Terms of use (n.d para 3.) [Internet]. [cited 2020 Jan 29]. Available from: http://www.phenoscanner.medschl.cam.ac.uk/about/

159. University of Manitoba. Concept Dictionary and Glossary for Population

Based Research (n.d. para. 1) [Internet]. [cited 2020 Jan 31]. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/resources/concept_dictionary.html

160. Wong KM, Langlais K, Tobias GS, Fletcher-Hoppe C, Krasnewich D, Leeds HS, et al. The dbGaP data browser: A new tool for browsing dbGaP controlled-access genomic data. Nucleic Acids Res. 2017;45(D1):D819–26.

161. Zylke JW, DeAngelis CD. Pediatric chronic diseases - Stealing childhood. J Am Med Assoc. 2007;297(24):2765–6.

162. Cleave J Van, Gortmaker SL, Perrin JM. Dynamics of Obesity and Chronic Health Conditions Among Children and Youth. JAMA - J Am Med Assoc. 2010;303.

163. Denny S, De Silva M, Fleming T, Clark T, Merry S, Ameratunga S, et al. The prevalence of chronic health conditions impacting on daily functioning and the association with emotional well-being among a national sample of high school students. J Adolesc Heal [Internet]. 2014;54(4):410–5. Available from: http://dx.doi.org/10.1016/j.jadohealth.2013.09.010

164. Compas BE, Jaser SS, Dunn MJ, Rodriguez EM. Coping with chronic illness in childhood and adolescence. Annu Rev Clin Psychol. 2012;8:455–80.

165. Perrin JM, Anderson LE, Van Cleave J. The rise in chronic conditions among infants, children, and youth can be met with continued health system innovations. Health Aff. 2014;33(12):2099–105.

166. Van Der Lee J, Mokkink L, Grootenhuis M, Heymans H, Offringa M. Definitions and Measurement of Chronic Health Conditions in Childhood A Systematic Review. JAMA. 2007;297(24):2741–51.

167. Toledano-Toledano F, Domínguez-Guedea MT. Psychosocial factors related with caregiver burden among families of children with chronic conditions. Biopsychosoc Med. 2019;13(1):1–9.

168. Shah R, Hagell A. International comparisons of health and wellbeing in

adolescence and early adulthood. 2019;(February):A2.3-A3.

169. Every Woman Every Child. Indicator and Monitoring Framework for the Global Strategy for Women's, Children's, and Adolescents' Health (2016-2030). 2017.

170. Barlow JH, Ellard DR. The psychosocial well-being of children with chronic disease, their parents and siblings: an overview of the research evidence base. Child Care Health Dev. 2006;32(1):19–31.

171. Mokkink LB, Van Der Lee JH, Grootenhuis MA, Offringa M, Heymans HSA. Defining chronic diseases and health conditions in childhood (0-18 years of age): National consensus in the Netherlands. Eur J Pediatr. 2008;167(12):1441–7.

172. Romano I, Buchan C, Baiocco-Romano L, Ferro MA. Physical-mental multimorbidity in children and youth: A scoping review. BMJ Open. 2021;11(5):1–9.

173. Romano I, Buchan MC, Ferro MA. Multimorbidity in children and youth: A scoping review protocol. BMJ Open. 2018;8(5):14–6.

174. Butler A, Van Lieshout RJ, Lipman EL, Macmillan HL, Gonzalez A, Gorter JW, et al. Mental disorder in children with physical conditions: A pilot study. BMJ Open. 2018;8(1):9–11.

175. Merikangas KR, Calkins ME, Burstein M, He JP, Chiavacci R, Lateef T, et al. Comorbidity of physical and mental disorders in the neurodevelopmental genomics cohort study. Pediatrics. 2015;135(4):e927–38.

176. Zahra A, Shang-Ming Z, Sinead B. Concept libraries for automatic electronic health record based phenotyping: A review. Int J Popul Data Sci. 2021;6(1):1–17.

177. Berry JG, Hall M, Cohen E, Feudtner C, Health C. Ways to Identify Children with Medical Complexity and the Importance of Why. 2015;229–37.

178. Hardelid P, Dattani N, Gilbert R. Estimating the prevalence of chronic

conditions in children who die in England, Scotland and Wales: A data linkage cohort study. BMJ Open. 2014;4(8):1–8.

179. Feudtner C, Hays RM, Haynes G, Geyer JR, Neff JM, Koepsell TD. Deaths attributed to pediatric complex chronic conditions: national trends and implications for supportive care services. Pediatrics. 2001;107(6).

180. Neff JM, Clifton H, Popalisky J, Zhou C. Stratification of children by medical complexity. Acad Pediatr [Internet]. 2015;15(2):191–6. Available from: http://dx.doi.org/10.1016/j.acap.2014.10.007

181. Agency for Healthcare Research and Quality. Chronic Condition Indicator (CCI) for ICD-9-CM [Internet]. [cited 2021 Dec 14]. Available from: https://hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp#overview

182. Agency for Healthcare Research and Quality, Healthcare Cost and Utilization project (HCUP). USER GUIDE : CHRONIC CONDITION INDICATOR FOR ICD-10-CM , (BETA VERSION). 2020;1(October).

183. Berry JG, Ash AS, Cohen E, Hasan F, Feudtner C, Hall M. Contributions of Children With Multiple Chronic Conditions to Pediatric Hospitalizations in the United States: A Retrospective Cohort Analysis. Hosp Pediatr. 2017;7(7):365–72.

184. Simon TD, Cawthon ML, Stanford S, Popalisky J, Lyons D, Woodcox P, et al. Pediatric medical complexity algorithm: A new method to stratify children by medical complexity. Pediatrics. 2014;133(6).

185. Simon TD, Cawthon ML, Popalisky J, Mangione-Smith R. Development and Validation of the Pediatric Medical Complexity Algorithm (PMCA) Version 2.0. Hosp Pediatr. 2017;7(7):373–7.

186. Simon TD, Haaland W, Hawley K, Lambka K, Mangione-Smith R. Development and Validation of the Pediatric Medical Complexity Algorithm (PMCA) Version 3.0. Acad Pediatr [Internet]. 2018;18(5):577–80. Available from: https://doi.org/10.1016/j.acap.2018.02.010

187. Feudtner C, Christakis DA, Connell FA. Pediatric deaths attributable to complex chronic conditions: a population-based study of Washington State, 1980-1997. Pediatrics. 2000;106(1 Pt 2):205-9.

188. Neff JM, Sharp VL, Muldoon J, Graham J, Popalisky J, Gay JC. Identifying and classifying children with chronic conditions using administrative data with the Clinical Risk Group classification system. Ambul Pediatr. 2002;2(1):71–9.

189. Feudtner C, Feinstein JA, Zhong W, Hall M, Dai D. Pediatric complex chronic conditions classification system version 2: Updated for ICD-10 and complex medical technology dependence and transplantation. BMC Pediatr. 2014;14(1):1–7.

190. Centers for Medicare & Medicaid Services. Health Care Code Sets: ICD-10 (MLN900943). 2021;(July):1–6.

191. Al Sallakh MA. Creating and Utilising the Wales Asthma Observatory to Support Health Policy, Health Service Planning and Clinical Research. Swansea; 2018.

192. Ferver K, Burton B, Jesilow P. The Use of Claims Data in Healthcare Research. Open Public Health J. 2009;2(1):11–24.

193. Hain, Devins. Directory of Life-Limiting conditions. 2014;3(February).

194. Fraser LK, Miller M, Hain R, Norman P, Aldridge J, McKinney PA, et al. Rising national prevalence of life-limiting conditions in children in England. Pediatrics. 2012;129(4).

195. Randall V, Cervenka J, Arday D, Hooper T, Hanson J. Prevalence of life-threatening conditions in children. Am J Hosp Palliat Med. 2011;28(5):310–5.

196. Chien AT, Kuhlthau KA, Toomey SL, Quinn JA, Houtrow AJ, Kuo DZ, et al. Development of the children with disabilities algorithm. Pediatrics. 2015;136(4):e871–8.

197. Royal College of Paediatrics and Child Health, University College London. Child Health Reviews- UK. Overview of child deaths in the four UK

countries. 2013;(September).

198. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. Second Edi. Waltham, MA: Morgan Kaufmann; 2012.

199. Patel D. Research data management: a conceptual framework. Libr Rev [Internet]. 2016;65(4–5):226–41. Available from: https://doi.org/10.1108/LR-01-2016-0001

200. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. PeerJ [Internet]. 2013;2013(1):1–25. Available from: https://doi.org/10.7717/peerj.175

201. Viseur R. Open science: Practical issues in open research data. DATA 2015 - 4th Int Conf Data Manag Technol Appl Proc [Internet]. 2015;201–6. Available from: https://doi.org/10.5220/0005558802010206

202. Childs S, McLeod J, Lomas E, Cook G. Opening research data: Issues and opportunities. Rec Manag J [Internet]. 2014;24(2):142–62. Available from: https://doi.org/10.1108/RMJ-01-2014-0005

203. Dai SQ, Li H, Xiong J, Ma J, Guo HQ, Xiao X, et al. Assessing the Extent and Impact of Online Data Sharing in Eddy Covariance Flux Research. J Geophys Res Biogeosciences [Internet]. 2018;123(1):129–37. Available from: https://doi.org/10.1002/2017JG004277

204. de Almeida UB, Fraga BMO, Giommi P, Sahakyan N, Gasparyan S, Brandt CH. Long-term multi-band and polarimetric view of Mkn 421: Motivations for an integrated open-data platform for blazar optical polarimetry. Galaxies [Internet]. 2017;5(4). Available from: https://doi.org/10.3390/galaxies5040090

205. Cragin MH, Palmer CL, Carlson JR, Witt M. Data sharing, small science and institutional repositories. Philos Trans R Soc A Math Phys Eng Sci [Internet]. 2010;368(1926):4023–38. Available from: https://doi.org/10.1098/rsta.2010.0165

206. Molloy JC. The open knowledge foundation: Open data means better science.

PLoS Biol [Internet]. 2011;9(12):1–4. Available from: https://doi.org/10.1371/journal.pbio.1001195

207. Ostell J. Data Sharing: Standards for Bioinformatic Cross-Talk. Hum Mutat [Internet]. 2009;30(4):vii–vii. Available from: https://doi.org/10.1002/humu.21013

208. Huang X, Hawkins BA, Lei F, Miller GL, Favret C, Zhang R, et al. Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conserv Lett [Internet]. 2012;5(5):399–406. Available from: https://doi.org/10.1111/j.1755-263X.2012.00259.x

209. Costello MJ. Motivating Online Publication of Data. Bioscience. 2009;59(5):418–27.

210. Parr CS. Open Sourcing Ecological Data. Bioscience [Internet]. 2007;57(4):309–10. Available from: https://doi.org/10.1641/b570402

211. Zuiderwijk A, Spiers H. Sharing and re-using open data: A case study of motivations in astrophysics. Int J Inf Manage [Internet]. 2019;49(May):228–41. Available from: https://doi.org/10.1016/j.ijinfomgt.2019.05.024

212. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PLoS One [Internet]. 2015;10(2):1–25. Available from: https://doi.org/10.1371/journal.pone.0118053

213. Lindemann N. What's the average survey response rate? [2021 benchmark] [Internet]. [cited 2022 Nov 14]. Available from: https://pointerpro.com/blog/average-survey-response-rate/#importance-of-response-rate-in-surveys

214. Ramshaw A. The Complete Guide to Acceptable Survey Response Rates [Internet]. [cited 2022 Nov 14]. Available from: https://www.genroe.com/blog/acceptable-survey-response-rate-2/11504#What-is-a-good-Survey-Response-rate

215. Chung L. What is a good survey response rate for online customer surveys?

[Internet]. [cited 2022 Nov 14]. Available from: https://delighted.com/blog/average-survey-response-rate#:~:text=So%2C what is a good,rate is 50%25 or higher.

11 APPENDIXES

11.1 APPENDIX 1: CONSENT FORM

Participant Identification Number for this research:

CONSENT FORM

Title of the research: Repeatable and reusable research - Exploring the needs of users for a Data Portal for Disease Phenotyping

Name of Researcher: Zahra Almowil

Name of Researcher's Supervisor: Prof. Sinead Brophy

Name of Researcher's Supervisor: Dr. Jodie Croxall

Please initial box

1. I confirm that I have read the information sheet for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐
2. 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. ☐
3. I understand that the information collected about me will be used to support other research in the future, and may be shared anonymously with other researchers. ☐
4. I agree to take part in the above study. ☐

Name of Participant	Date	Signature
Name of Person taking consent	Date	Signature

11.2 APPENDIX 2: PARTICIPANT INFORMATION SHEET

Participant Information Sheet

Study title

Repeatable and reusable research - Exploring the needs of users for a Data Portal for Disease Phenotyping

Invitation and brief summary

You are being invited to take part in this interview for the research project as part of my PhD research degree in Health Informatics at the Data Science Building, Swansea University Medical School. Before you decide to participate in this interview, it is highly recommended to understand the purpose of the research and what it will involve. I really appreciate your time in reading the following information, and please feel free to ask us about anything that is not clear or if you need more information. Thank you for your time in this matter.

What is the purpose of the research project?

The purpose of this research project is to inform the design of the Wales Portal for Data through development and evaluation of interviews, focus groups, and surveys with key stakeholders to improve disease phenotyping.

Do I have to take part?

No, you do not have to participate in this research project. Participation is totally voluntary, and you have the right to withdraw at any time without giving a reason and without prejudice.

What will my involvement require?

If you agree to participate in this interview, please read this sheet, tick the box to indicate that you have read this participant information sheet, and sign a consent form.

What are the possible benefits of taking part?

Your contribution in this research is invaluable because it will be used to inform the development of a Portal to improve access to, usability, and collaborations with routine data.

Will my taking part in the study be kept confidential?

Yes, all the information collected from you during the interview will be anonymous. For further protection of your privacy and confidentiality, all records will be destroyed after transcriptions, and

transcription is password protected in a secure server.

Who is organizing and funding the research?

This research project which is part of my PhD research in Health Informatics, is organized and supervised by my supervisors who are working with the Health Data Research UK Wales and Northern Ireland Site, Swansea University Medical School, and is funded by Kuwait Cultural Office in London.

Further information and contact details:

If you have any questions concerning this research, please contact:

Zahra Almowil (Researcher):

Prof. Sinead Brophy (Researcher's Supervisor):

Dr. Jodie Croxall (Researcher's Supervisor):

Thank you for taking the time to read this Information Sheet.

Prifysgol Abertawe
Swansea University
Ysgol Feddygaeth
Medical School

Application for Standard Ethical Approval

CHECKLIST

Please note that we are able to review an application only when all documentation is submitted alongside this application form. Should any necessary appendices not be attached, this could delay the submission until the following month. Please use this checklist below to ensure that the application is complete. Many thanks.

	Attached Yes / No / N/A	Comment
Recruitment advertisement or email(s)	N/A	
Participant information sheet(s)	Yes	
Consent form(s)	Yes	
Debrief sheet(s)	N/A	
Questionnaire(s)	No	
Interview or Focus Group schedule(s)/questions	Yes	
Workshop schedule(s)/questions	N/A	

Written consent from public or private body	N/A	
Supervisor signature	Yes	

1. PLEASE COMPLETE THE FORM USING TYPESCRIPT

2. (hand-written applications will not be considered)

Principal Investigator	Zahra Ahmed Almowil
Date	21/01/19
School	Swansea University
E-mail address	[redacted]
Title of Proposed Research	Development and Evaluation of a Secure Data Portal and Interactive Platform of Disease Phenotyping for Retrieval and Analysis of the Terminologies of Diagnoses, Symptoms, Medications and Procedures in Wales
Type of Researcher (Please tick)	Postgraduate student
Name of course & supervisor	Name of course: PhD in Health Informatics Dr Shangming Zhou

	Prof. Sinead Brophy
Supervisor e-mail address	[redacted]
Qualifications and professional background	Master Degree in Heath Information Systems Clinical Instructor in the Medical Records Department at the College of Heath Sciences, Kuwait for the last 13 years.

3. Briefly describe the rationale and the main aims of the research you wish to undertake, including a statement of the intended benefits of the research. Please use non-technical language wherever possible.

Aims:

To inform the design of the Wales Portal for Data through development and evaluation of interviews, focus groups and surveys with key stakeholders to improve disease phenotyping.

Objectives:

The project focuses on:

- Identifying the associated literatures and reports for the disease phenotypes focusing on chronic conditions - asthma, dementia, diabetes, cancers.
- Design and development of the platform;
- Identifying the phenotypes of diseases in adults from the medical codes;
- Performing a needs assessment (e.g., SWOT analysis of the current situation and what is needed for the future), a scoping study of what resources are available to deliver the Portal, and examining user experience and recommendations.
- Developing methods of evaluating the prototype of the portal to meet the demands

The intended benefits of the research:

This work will greatly inform the development of a Portal to improve access to and

collaborations with routine data. This is part of an all-UK agenda and the finding of this work will have implications for other countries working to access and share routine such as in Australia and Kuwait.

4. Briefly describe the overall design of the project including dates and/or the proposed period of investigation

Contact with experts in the field about the state-of-the-art research in portal development from May 2019 to July 2019

5. Briefly describe the methods of data collection and analysis. Please describe all measures to be employed. If questionnaire or interviews are to be used, please provide the questionnaire / interview questions and schedule.

Materials and methods:

This project consists of the following stages:

1) Scoping through literature review to examine the definitions of diseases in adults, focusing on chronic conditions - asthma, dementia, diabetes, cancers.

2) Scoping through literature review and contact with experts in the field to examine what Portals already exist, how are they designed, how are they used, how we might build from the current situation.

3) Needs assessment – a focus group with researchers (n=5-8) working with SAIL data within the Data Science Building, Swansea University, will be held for 2 hours to perform a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system and the proposed Portal. Interviews with researchers working with SAIL data outside the Data Science Building will be held to take their opinions using the same SWOT analysis approach. The focus group and interviews will all be transcribed verbatim and analysed with recommendations and learning written up to be submitted to Health Informatics journals or the International Journal of

Population Data Science.

4) Scoping of resources and requirements – interviews with the health informatics teams in both Swansea and other universities to examine what is needed to develop common metadata standards, sharing (e.g., Via WIKI?) of information, how to run a centralised or federated approach to the Portal, show to deal with support and referencing of codes and terms, how to host information, user interface and knowledge of what works best, how to make it very easy to use with a low barrier to entry. These findings from this work will also be submitted to Health Informatics Journals.

5) Recommendations and requests from policy makers in terms of visualization of data.

6) Evaluation and interpretation of the prototype portal – examining user experience and recommendations through (a) one to one interview with researchers as they use the Portal (2) online survey using survey monkey for all users as it is rolled out. Multi-criteria decision making will be used to combine the recommendations and yield an overall recommendation that will be fed back to the developers.

6. Location of the proposed research (i.e., Departmental labs, schools, etc)

Data Science Building, Swansea University

7. Describe the participants: give the age range, gender, inclusion and exclusion criteria, and any particular characteristics pertinent to the research project.

Researchers, Health Informatics, and Policy makers

Age 18+, M+F, work in the field of big data

8. How will the participants be selected and recruited? Please describe in detail the process of recruitment, including how and by whom initial contact is made with participants (e.g., advertisement, e-mail).
I will inform the participants orally through HDRUK (Health Data Research UK) contacts.

9. What procedures (e.g., interviews, computer-based learning tasks, etc.) will be used to gather information from participants?
Interviews and focus group

10. What potential risks to the participants do you foresee and how do you propose to ameliorate/deal with potential risks? Declare any relationship with the participants.
No personal identifiable information will be collected. No foreseen potential risk identified so far.
11. What potential risks to the interests of the researchers do you foresee and how will you ameliorate/deal with potential risks?
No foreseen potential risks identified so far.

12. How will you brief and debrief participants? (Please attach copy of participant information sheets and relevant debrief information)
I will summarize the discussion at the end of the session to confirm summary is representative of participants opinions

13. Will informed consent be sought from participants?	Yes (Please attach a copy of the consent form and participant information sheet)	Yes

14. If there are doubts about participants' abilities to give informed consent, what steps have you taken to ensure that they are willing/competent to participate?
N/A

15. If participants are under 18 years of age, please describe how you will seek informed consent.
N/A

16. How will consent be recorded?
Paper based written signature

17. Will participants be informed of the right to withdraw from your study without penalty? If no, please explain why.
Yes

18. How do you propose to ensure participants' confidentiality and anonymity?

No names will be collected from the participants. They will be given numbers and referred to in terms of numbers

19. Please describe the arrangements for storing and disposal of data:

a). Please describe the arrangements for storing and disposal of data:

Record and transcribe, and records will be destroyed after transcriptions.

b). Please explain, for each of the above, the arrangements you will make for the security of the data

Transcription is password protected in a secure server.

20. Does your research require the written consent of a public or private body, e.g., school, local authority or company? If so, please attach letter of consent.

No

21. If your proposed research is with 'vulnerable' groups (e.g., children, people with a disability etc.), has an up-to-date Disclosure and Barring Service (DBS) check (previously CRB check) if UK, or equivalent non-UK clearance been requested and/or obtained for all researchers?

N/A

22. Does your research involve the collection of Human Tissue? E.g., saliva, urine	Yes	
	No	x

Applicant's signature: _____ Date: _____

Supervisor's signature: _____ Date: _____

(If appropriate)

Upon completion, please forward an electronic copy (as a single document, Word or PDF) by e-mail to sumsresc@swansea.ac.uk and a signed hard copy to the Chair of the Committee, Dr Deyarina Gonzalez.

Administrative Support

Research Ethics Sub- Committee,

SUMS

Swansea University

Singleton Park, Swansea, SA2 8PP.

Dr Deyarina Gonzalez

Research Ethics Sub-Committee,

SUMS

Swansea University

Singleton Park, Swansea, SA2 8PP.

Email: [redacted]

Chairperson REG

****RESEARCH MAY ONLY COMMENCE ONCE ETHICAL**

APPROVAL HAS BEEN OBTAINED**

11.4 APPENDIX 4: ETHICAL APPROVAL

Ethical Approval

Ethics Committee Use Only

Principal Investigator	Zahra Ahmed ALMOWIL
Title of Proposed Research	Development and Evaluation of a Secure data Portal and Interactive Platform of Disease Phenotyping for Retrieval and Analysis of the Terminologies of Diagnoses, Symptoms, Medications and Procedures in Wales.
RESC Project reference number	2019-0007

Application approved	**Yes**	X	**No**			
Conflict of interest	**Yes**		**No**	X		
If yes, please supply details						

Chair of SUMS RESC	Deya Gonzalez Associate Professor of Molecular Medicine [redacted] Swansea University Medical School Singleton Park, Swansea, SA2 8PP, UK. Email [redacted] [redacted]
Date 07.03.19	Signature [redacted]

This application **has been granted ethical approval** in its current form.

Please ensure that you quote project reference number 2019-0007 in any correspondence with the SUMS RESC

Time limit for applicant to respond	(Two months from receipt of email from ethics panel)

11.5 Appendix 5: Application for Standard Ethical Approval (2)

Application for Standard Ethical Approval

CHECKLIST

Please note that we are able to review an application only when all documentation is submitted alongside this application form. Should any necessary appendices not be attached, this could delay the submission until the following month. Please use this checklist below to ensure that the application is complete. Many thanks.

	Attached Yes / No / N/A	Comment
Recruitment advertisement or email(s)	N/A	
Participant information sheet(s)	Yes	
Consent form(s)	Yes	
Debrief sheet(s)	N/A	

Questionnaire(s)	Yes	See Appendix .3
Interview or Focus Group schedule(s)/questions	Yes	
Workshop schedule(s)/questions	N/A	
Written consent from public or private body	N/A	
Supervisor signature	Yes	

- **PLEASE COMPLETE THE FORM USING TYPESCRIPT**
- **(hand-written applications will not be considered)**

Principal Investigator	Zahra Ahmed Almowil
Date	15/02/21
School	Swansea University
E-mail address	[redacted]
Title of Proposed Research	Repeatable and reusable research - Exploring the needs of users for a Data Portal for Disease Phenotyping

Type of Researcher (Please tick)	Postgraduate student (PhD)
Name of course & supervisor	Name of course: PhD in Medical and Health Care Studies - Health Informatics Prof. Sinead Brophy Dr. Jodie Croxall
Supervisor e-mail address	[redacted] [redacted]
Qualifications and professional background	Master Degree in Heath Information Systems Clinical Instructor in the Medical Records Department at the College of Heath Sciences, Kuwait for the last 13 years.

23. Briefly describe the rationale and the main aims of the research you wish to undertake, including a statement of the intended benefits of the research. Please use non-technical language wherever possible.

Aims:

To inform the design of the Wales Portal for Data through development and evaluation of interviews, focus groups and surveys with key stakeholders to improve disease phenotyping.

Objectives:

The project focuses on:

- Identifying the associated literatures and reports for the disease phenotypes focusing on chronic conditions - asthma, dementia, diabetes, cancers.
- Design and development of the platform;
- Identifying the phenotypes of diseases in adults from the medical codes;
- Performing a needs assessment (e.g., SWOT analysis of the current situation and what is needed for the future), a scoping study of what resources are available to deliver the Portal, and examining user experience and recommendations.
- Developing methods of evaluating the prototype of the portal to meet the demands

The intended benefits of the research:

This work will greatly inform the development of a Portal to improve access to and collaborations with routine data. This is part of an all-UK agenda and the finding of this work will have implications for other countries working to access and share routine such as in Australia and Kuwait.

24. Briefly describe the overall design of the project including dates and/or the proposed period of investigation

Research Plan:

- **October 2018~April 2019**

Candidature report and scoping through literature review to examine the definitions of diseases in adults

- Requesting ethical approval from the SUMS Research Ethics Sub-Committee and training in qualitative methods

- **May 2019 ~ July 2019**

Scoping through literature review and contact with experts in the field about the state-of-the-art research in portal development

- **August 2019 ~ January 2020**

Needs assessment of focus groups and key stakeholders

- **February 2020 ~ June 2020**

Scoping of resources and requirements

- **July 2020 ~ February 2021**

Developing the methods to evaluating the prototype portal.

- **March 2021 ~ September 2021**

Write the thesis and submit it for viva.

25. Briefly describe the methods of data collection and analysis. Please describe all measures to be employed. If questionnaire or interviews are to be used, please provide the questionnaire / interview questions and schedule.

Materials and methods:

<u>This project consists of the following stages:</u>

1) Scoping through literature review to examine the definitions of diseases in adults, focusing on chronic conditions - asthma, dementia, diabetes, cancers.

2) Scoping through literature review and contact with experts in the field to examine what Portals already exist, how are they designed, how are they used, how we might build from the current situation.

3) Needs assessment – a focus group with researchers (n=5-8) working with SAIL data within the Data Science Building, Swansea University, will be held for 2 hours to perform a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system and the proposed Portal. Interviews with researchers working with SAIL data outside the Data Science Building will be held to take their opinions using the same SWOT analysis approach. The focus group and interviews will all be transcribed verbatim and analysed with recommendations and learning written up to be submitted to Health Informatics journals or the International Journal of Population Data Science.

4) Scoping of resources and requirements – interviews with the health informatics teams in both Swansea and other universities to examine what is needed to develop common metadata standards, sharing (e.g., Via WIKI?) of information, how to run a centralised or federated approach to the Portal, show to deal with support and referencing of codes and terms, how to host information, user interface and knowledge of what works best, how to make it very easy to use with a low barrier to entry. These findings from this work will also be submitted to Health Informatics Journals.

5) Recommendations and requests from policy makers in terms of visualization of data.

6) Evaluation and interpretation of the prototype portal – examining user experience and recommendations through (a) one to one interview with researchers as they use the Portal (2) online survey using survey monkey for all users as it is rolled out. Multi-criteria decision making will be used to combine the recommendations and yield an overall recommendation that will be fed back to the developers.

26. Location of the proposed research (i.e., Departmental labs, schools, etc)

Data Science Building, Swansea University

27. Describe the participants: give the age range, gender, inclusion and exclusion criteria, and any particular characteristics pertinent to the research project.

Researchers, Health Informatics, and Policy makers

Age 18+, M+F, work in the field of big data

28. How will the participants be selected and recruited? Please describe in detail the process of recruitment, including how and by whom initial contact is made with participants (e.g., advertisement, e-mail).

I will inform the participants orally through Health Data Research UK (HDRUK) contacts. I will invite researchers who have validated disease concepts to take part in an interview. I will email them and I will phone them to ask for their participations. I will email researchers to ask them to complete an online questionnaire.

29. What procedures (e.g., interviews, computer-based learning tasks, etc.) will be used to gather information from participants?
Interviews, focus group, and questionnaires

30. What potential risks to the participants do you foresee and how do you propose to ameliorate/deal with potential risks? No personal identifiable information will be collected. No foreseen potential risk identified so far.
31. What potential risks to the interests of the researchers do you foresee and how will you ameliorate/deal with potential risks?
No foreseen potential risks identified so far.

32. How will you brief and debrief participants? (Please attach copy of participant information sheets and relevant debrief information)
I will summarize the discussion at the end of the session to confirm summary is representative of participants opinions

33. Will informed consent be sought from participants?	Yes (Please attach a copy of the consent form and participant information sheet)	Yes

34. If there are doubts about participants' abilities to give informed consent, what steps have you taken to ensure that they are willing/competent to participate?
N/A

35. If participants are under 18 years of age, please describe how you will seek informed consent.
N/A

36. How will consent be recorded?
Paper based written signature

37. Will participants be informed of the right to withdraw from your study without penalty? If no, please explain why.
Yes

38. How do you propose to ensure participants' confidentiality and anonymity?
No names will be collected from the participants. They will be given numbers and referred to in terms of numbers

39. Please describe the arrangements for storing and disposal of data:
a). Please describe the arrangements for storing and disposal of data: Record and transcribe, and records will be destroyed after transcriptions. b). Please explain, for each of the above, the arrangements you will make for the security of the data

Transcription is password protected in a secure server. Interviews will be tape recorded, and the recording will be transcribed. Once transcripts are electronic, records will be deleted. All transcripts will be stored in a locked area and password protected and computers are password protected also.

40. Does your research require the written consent of a public or private body, e.g., school, local authority or company? If so, please attach letter of consent.
No

41. If your proposed research is with 'vulnerable' groups (e.g., children, people with a disability etc.), has an up-to-date Disclosure and Barring Service (DBS) check (previously CRB check) if UK, or equivalent non-UK clearance been requested and/or obtained for all researchers?
N/A

42. Does your research involve the collection of Human Tissue? E.g., saliva, urine	Yes	
	No	x

Applicant's signature: Zahra Date: 15/02/2021

Supervisor's signature: Sinead Date: 15/02/2021

(If appropriate)

Appendix 1. Participant Information Sheet

Participant Information Sheet

Study title

Repeatable and reusable research - Exploring the needs of users for a Data Portal for Disease Phenotyping

Invitation and brief summary

You are being invited to take part in this interview for the research project as part of my PhD research degree in Health Informatics at the Data Science Building, Swansea University Medical School. Before you decide to participate in this interview, it is highly recommended to understand the purpose of the research and what it will involve. I really appreciate your time in reading the following information, and please feel free to ask us about anything that is not clear or if you need more information. Thank you for your time in this matter.

What is the purpose of the research project?

The purpose of this research project is to inform the design of the Wales Portal.

Do I have to take part?

No, you do not have to participate in this research project. Participation is totally voluntary, and you have the right to withdraw at any time without giving a reason and without prejudice.

What will my involvement require?

If you agree to participate in this interview, please read this sheet, tick the box to indicate that you have read this participant information sheet, and sign a consent form.

What are the possible benefits of taking part?

Your contribution in this research is invaluable because it will be used to inform the development of a Portal to improve access to, usability, and collaborations with routine data.

Will my taking part in the study be kept confidential?

Yes, all the information collected from you during the interview will be anonymous. For further protection of your privacy and confidentiality, all records will be destroyed after transcriptions, and transcription is password protected in a secure server.

Who is organizing and funding the research?

This research project which is part of my PhD research in Health Informatics is organized and supervised by my supervisors who are working with the Health Data Research UK Wales and Northern Ireland Site, Swansea University Medical School, and is funded by Kuwait Cultural Office in London.

Further information and contact details:

If you have any questions concerning this research, please contact:

Zahra Almowil (Researcher): [redacted]

Prof. Sinead Brophy (Researcher's Supervisor): [redacted]

Dr. Jodie Croxall (Researcher's Supervisor): [redacted]

Thank you for taking the time to read this Information Sheet.

Appendix 2. Consent Form

Participant Identification Number for this research:

CONSENT FORM

Title of the research: Repeatable and reusable research - Exploring the needs of users for a Data Portal for Disease Phenotyping

Name of Researcher: Zahra Almowil

Name of Researcher's Supervisor: Prof. Sinead Brophy

Name of Researcher's Supervisor: Dr. Jodie Croxall

Please initial box

1. I confirm that I have read the information sheet for the above research. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. 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. ☐

3. I understand that the information collected about me will be used to support other research in the future, and may be shared anonymously with other researchers. ☐

4. I agree to take part in the above study. ☐

_____	_____	_____
Name of Participant	Date	Signature

_____	_____	_____

Name of Person taking consent	Date	Signature

11.6 Appendix 6: The Concept Library Usability Scale

* 1. I think that I would like to use this concept library frequently.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Could you please explain why you disagreed or agreed with this statement?

* 2. I think that I would need the support of a technical person to be able to use this concept library.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

Could you please explain why you disagreed or agreed with this statement?

* 3. I found the various functions in this concept library, such as searching and viewing concepts; and creating and editing concepts, were easy to use.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 4. I felt very confident using the concept library.

- Strongly disagree
- Disagree
- Neither agree nor disagree

- Agree
- Strongly agree

* 5. I needed to learn a lot of things before I could get going with this concept library.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 6. I think that the user documentation is task oriented and consists of clear, step by step instructions.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 7. I found the concept library supports advanced functional tasks (e.g., it allows using of programming languages such as R, SQL, or Python).

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 8. I feel it is acceptable if I am required to reference the concept library when publishing papers.

- Strongly disagree
- Disagree
- Neither agree nor disagree

- Agree
- Strongly agree

* 9. I think the concept library is interoperable with other required/ related systems.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 10. I found that it was easy to understand how the concept library is run and managed.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 11. I thought the concept library supports clear algorithms labeling convention.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

* 12. Can you tell us more about why you give the answers you did? (e.g., what could be improved? and what did you like about the system?

13. If you are interested in participating in one-to-one interview, please provide us with your contact information.

Name

Company

Address

Address 2

City/Town

County

Post Code

Country

Email Address

Phone Number

Thank you for completing the Survey.