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

Zahra Almowil, Shang-Ming Zhou, Sinead Brophy, Jodie Croxall

Submitted to: JMIR Human Factors
on: June 07, 2021

Table of Contents

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

Zahra Almowil[1] MSc; Shang-Ming Zhou[2] PhD; Sinead Brophy[3] PhD; Jodie Croxall[4] PhD

[1]Swansea University Swansea, Wales GB
[2]Centre For Health Technology, Faculty Of Health, University Of Plymouth, Plymouth, PL4 8AA, UK Plymouth GB
[3]Swansea University Medical School, Wales SA2 8PP Swansea, Wales GB
[4]Biomedical Sciences, Swansea University, Wales SA2 8PP Swansea, Wales GB

Corresponding Author:
Zahra Almowil MSc
Swansea University
Data Science Building
Swansea University
Swansea, Wales
GB

Abstract

Background: Big data research in the health field is hindered by a lack of agreement in how to identify and define different disease conditions and their medications. This means researchers and health professionals often have different definitions of the same condition. This lack of agreement makes it difficult to compare different study findings and so hinders the field's ability to do repeatable and reusable research.
Aim: The aim of this study was to examine the views and needs of: 1) users including researchers, health professionals, and clinicians, and 2) designers such as the health informatics teams, in creating a portal of definitions for disease phenotyping (a concept library).

Objective: The aim of this study was to examine the views and needs of: 1) users including researchers, health professionals, and clinicians, and 2) designers such as the health informatics teams, in creating a portal of definitions for disease phenotyping (a concept library).

Methods: Qualitative study using interviews and a focus group. One to one interview with researchers, clinicians and managers have been conducted (n=6) to examine their specific needs. In addition, a focus group with participants (n=14) working with the SAIL databank, a national e-health data linkage infrastructure, was held to perform a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system and the proposed concept library. The interviews and the focus group were analysed separately following Braun and Clarkes (2006) analysis approach.

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

Conclusions: The study indicated that most interviewees would value a concept library for disease phenotyping. However, only half of the participants felt they would contribute to 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 revealed that different stakeholders have different requirements, facilitators, barriers, and concerns of a prototype concept library.

(JMIR Preprints 07/06/2021:31021)
DOI: https://doi.org/10.2196/preprints.31021

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

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

Zahra Ahmed Almowil, Shang-Ming Zhou, Jodie Croxall, Sinead Brophy.

Zahra Almowil, M.Sc., School of Medicine, Swansea University, Wales SA2 8PP
Shangming Zhou, Senior Lecturer in Data Science, Swansea University, Wales SA2 8PP
Jodie Croxall, Associate Professor of Biomedical Sciences, Swansea University, Wales SA2 8PP
Sinead Brophy, Professor of Public Health Data Science, Swansea University, Wales SA2 8PP

Corresponding Author:
Zahra Almowil, M.Sc.
School of Medicine, Swansea University
Wales SA2 8PP
The UK
Phone: 07552894384
Email: 934467@swansea.ac.uk

Abstract

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

Objective: 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: A qualitative study using interviews and a focus group. 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 with researchers (n=14) working with the SAIL databank, a national e-health data linkage infrastructure, was held to perform a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system for phenotyping and the proposed concept library. The interviews and focus group were both verbatim transcribed, and two 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 many requirements needed before its development. Although all the participants stated that they were aware of some existing concept libraries, the majority of them expressed negative perceptions about them. The participants mentioned several facilitators that would stimulate them to share their work and/or to reuse the work of others, and they pointed out several barriers that could inhibit them from sharing their work and/or reusing the work of others. The participants have suggested some developments they would like to see to improve reproducible research output using routine data.
Conclusions: The study indicated that most interviewees would value a concept library for phenotypes. However, only half of the participants felt 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 revealed that different stakeholders have different requirements, facilitators, barriers, and concerns about a prototype concept library.

KEYWORDS
Electronic Health Records; Record linkage; Reproducible research; Clinical codes; Concept libraries

Introduction

Health care systems are becoming more digitally focused rather than paper based, and are moving to Electronic Health Records (EHRs) [1]. This means there is the availability of large amounts 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 this data effectively [2-5]. The growing availability of electronic patient data offers health care practitioners increased opportunities for secondary use of EHRs data to improve quality of care and research [6-8]. However, present literature does not describe the barriers that make data utilisation and deidentification processes difficult, nor does it focus on users' practical needs for data linking [9]. “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 UK, which gathers data about general practise from the following databases [13]: Clinical Practice Research Data Link (CPRD) [14], The Health Improvement Network (THIN) [15], QResearch [16], and Secured Anonymized Information Linkage (SAIL) [17]. However, with different datasets (e.g., hospital, general practise, emergency care), defining a condition is still very subjective, as there are many phenotyping algorithms for identifying the same condition (e.g., there are currently 66 ways of defining asthma using routine health data) [18], and interpretation or manipulation of data often requires knowledge of complex programming languages, such as SQL [4]. This means that EHRs are still not really accessible to many, as their use requires specialized programming skills.

One of the most important objects 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 [8,19]. 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 the patients with or without conditions [19], and researchers would not be able to compare studies effectively. Even though code lists are available in some research, researchers often encounter difficulties 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, causing 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 EHRs-driven studies [22].

Although researchers request better transparency in sharing clinical code lists [23,24], they face difficulties in obtaining comprehensive code lists from EHRs-based research. While there are currently no obligations form journals and funding parties to publish code lists, the STROBE and RECORD initiatives encourage transparency and open access to publicly available EHRs-based research [25-27]. To address the 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 ClinicalCodes.org [22], CALIBER data portal [4], and the Concept Dictionary at the Manitoba Centre for Health Policy [28]. Building online 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 validation of EHRs-based research [22]. The purpose of this study was to explore the needs of various users, including researchers, clinicians, machine learning experts, and managers, for the development of a data portal for phenotypes (a concept library), and to examine why existing concept libraries are not more widely used.

Methods

Design

A qualitative study using one-to-one interviews and a focus group. We recruited a small purposive sample for in-depth one-to-one interviews in the first phase because they allow 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 bigger sample of participants for the focus group in order 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 the purpose of this research, we adopted a semi-structured approach. We created semi-structured interview questions based on the Krueger and Casey format [29], which include introductory, flow, key, and final questions to be used in one-to-one interviews (presented in table 1). Also, we created a list of ten questions based on the objectives of this research for the focus group session. The purpose of the questions was to generate thoughtful and thorough responses from the participants, therefore closed-ended questions (e.g., yes or no) were avoided. Both the interviews and the focus group were audio recorded and transcribed verbatim, and two thematic analyses were performed, using the six steps of Braun and Clarke to identify themes and subthemes [30].

Table 1. One-to-one interviews questions guide

Introductory questions	Follow questions	Key questions	Final questions
In order to improve repeatable research in Swansea, a team of developers is developing a prototype concept library. This is a portal, which allows access to the READ codes or ICD10 codes to identify conditions. Do you think this will be a helpful resource? Is	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 in order to modify them? In your opinion, how should codes and algorithms be validated, and should they be validated? (Why should/should not?) There are often	What are your requirements for the concept library in order for it to be helpful and user-friendly? What developments would you like to see to improve repeatable research using routine data?

the concept library a good idea that we should continue to develop?	different versions of 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, suspected)? or do you think we should put all possible methods of identifying a condition, valid or not and allow the researcher to choose?	

Data Collection

The first author asked six participants from a variety of disciplines, including researchers (n=3), a clinician (n=1), a machine learning expert (n=1), and a senior research manager (n=1) at Swansea University and Cardiff University to participate in one-to-one interviews by email. The invitation email specified the aim and the 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 Krueger and Casey's structure [29], were used (presented in table 1). The structure of the interview questions consists 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 the authors intended to explore. The final questions were designed to summarize the interview and to 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 (e.g., their office). After 5 interviews, no new themes were emerging and interview 6 confirmed that no new themes emerged. The interviews were audio recorded and were transcribed verbatim. Then, thematic analyses were performed, using the six steps of Braun and Clarke to identify themes and subthemes [30].

All researchers working with the SAIL databank, a national e-health data linkage infrastructure in Wale (n=34) were invited by email to participate in the focus group, and fourteen researchers attended the focus group. Two focus group discussions, which each had seven participants, were held for two hours by two moderators (ZA and SB), who used the same set of semi-structured questions to perform a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) for the current system for phenotyping and the proposed concept library. We used a SWOT analysis tool in

this research because it enabled the participants to discuss what they like (Strengths), what advantages would be gained (opportunities) and what problems (Weaknesses) and issues (Threats) they felt needed to be tackled. Although the two moderators utilised the same set of questions, the order of the questions was adjustable to the needs of each group.

At the beginning of the focus group, the first author conducted 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 UK. Then a second presentation about the Swansea University prototype concept library was 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 strength 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've talked about, what is most important to you?
- Is there anything we should have talked about, but didn't?

The goal of employing the SWOT analysis was to identify positive factors that operate together as well as potential difficulties that must be identified and may be 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. Figure 1 presents a summary of identified strengths, weaknesses, opportunities, and threats in the current system for phenotyping and the proposed concept library. The two focus group discussions were audio recorded and were transcribed verbatim. Then, thematic analyses were conducted using the six steps of Braun and Clarke to discover the main themes and sub themes (see table 3).

Figure 1: 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 amongst 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).

Data Analysis

The interviews and the focus group were analysed separately following Braun and Clarke's (2006) analysis approach [30]. The transcripts of the interviews and the focus group were read several times and then initial codes were grouped into themes and subthemes using a qualitative data analysis software (NVivo) [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 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. They critically reviewed themes again to determine their primary meanings, and similar initial themes were joined into one theme. They discussed the definitions of the relevant themes to the research questions and applied appropriate names to describe each in this article. See table 2 for a further description of the thematic analytic steps.

Table 2. The six thematic analytic steps used for this research

1. Self-Familiarizing with the data ZA 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 were transcribed by professional transcribers. During this phase, ZA read all of the interview and focus group 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 organised them in a note book.	2. Creating initial codes After familiarising 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. They organised the identified codes into meaningful groups using qualitative data analysis software (NVivo). They used the same coding procedure for all the transcripts.
3. 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. Then, using the NVivo software, the initial codes were sorted and labelled into themes and subthemes depending on the	4. 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

meaning or relations shared by the codes.	to the research question.
5. 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 article. 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.	6. 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.

Results

Interviews with users:

Six one-to-one interviews were conducted, and each interview lasted for about half an hour. The analysis of the interviews resulted in four main themes, with several subthemes (presented in table 3). The four main themes were:

1) Prior 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 3. Presentation of the themes and subthemes of one-to-one interviews

Themes	Examples of participant narratives
<u>Theme (1) Prior opinion of a prototype concept library</u>	
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	
1. 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).
2. 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).

1) Prior opinion of a prototype concept library

The majority of the participants were positive about the prototype concept library and felt 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 ICD 10 codes from hospitals are the most common data items that machine learning experts would like to use the majority of the time. They use data linkage repositories to extract the necessary data for machine learning 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, two 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 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).

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 it needs to be simple and easy to use by naive users. "*It should be simple enough, within one or two clicks; we can find the required data*" (researcher 3), but also should contain advanced expert features (R, SQL or Python programming languages) to extract, include, or exclude codes necessary for their studies. *"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). Also, they stated that the concept library should have a good search engine so that they could easily find the phenotypes and phenotyping algorithms they wanted 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 (for example, are codes from general practice, hospital (ICD, SNOMED), BNF medication etc.), 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. *"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."* (a clinician), 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, 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 about how to validate phenotyping algorithms, and the majority of 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 needed to 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.

3) Experience of existing concept libraries

All the participants stated that they were aware of some existing concept libraries, such as CALIEBER and ClinicalCodes.Org (both in the UK), but the majority of them did not use them. The reasons given for not using them included: they already have their own self-made concept libraries (e.g., concepts they have used before) or because the available concept libraries do not provide phenotyping algorithms that fit their studies. For example, a machine learning expert mentioned the reasons of not using two of the existing concept libraries, which are the Concept Dictionary at the Manitoba Centre for Health Policy in Canada, and CALIEBER in the UK were *"Canadian systems provide Canadian data for their studies, CALIEBER is specific for cardiovascular disease and does not have many concepts in it"*. Conversely, two of the participants mentioned that they have 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).

4) Recommendations to improve repeatable research

The particpants 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), and 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).

The focus group:

Out of the 34 invited researchers, 14 people attended the focus group. These participants were researchers (n=14) from Swansea University who were working with the SAIL data in the Data Science Building. There were five female participants and nine male participants out of a total of fourteen. Six of the 14 participants were PhD holders, six were Masters holders, and two were Bachelor's degree holders (see table 4).

Table 4. A summary of general information on the participants in the focus groups

Current job position	13 Data Scientist 1 Financial Planner
Sex	5 Females (36%) 9 Males (64%)
Education	6 PhD degree (43%) 6 Master degree (43%) 2 Bachelor degree (14%)
Research interests	1. 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

2. Financial Planners
    - Intervention between primary care and secondary care and how they interact

The focus group was held for 2 hours to perform a SWOT analysis (Strengths, Weaknesses, Opportunities, Threats) of the current system for phenotyping and the proposed concept library, the focus group 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 and barriers to participants' contributing their research methods
2) Facilitators and barriers to participants' usage of other researchers' research 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' usage and perceptions of existing concept libraries

1) Facilitators and barriers to participants' contributing their research methods

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). "*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 organise team resources and discuss research findings from other researchers who used their code lists. However, improving research opportunities, increasing academic achievement, and sharing of 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 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 (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., lots of clinical researchers have created phenotyping algorithms for particular research, and these algorithms are hard to be generalized).

Several participants mentioned that lack of return for their hard work (e.g., not getting 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).

2) Facilitators and barriers to participants' usage of other researchers' research methods

Facilitators:

The participants mentioned several facilitators that would encourage them to reuse research methods developed by others, such as: 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), and 3) using work of others as a reference to compare research outcome, and researchers want to prove that there is a basis for the use of such codes.

Barriers:

Conversely, the participants pointed out several barriers which could inhibit them from reusing methods developed by other researchers: 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 have no relevance to finding conditions in GP data in the UK. *"Yes, it works in their population, because where they've trained it* " (data scientist 5), and 3) whether the data is 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).

3) The 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 concern about the completeness rate of phenotyping algorithms. They would like to know the percent 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). Also, there has been a question as to whether codes need to be peer reviewed so that quality is reviewed.

4) The requirements of the participants for the prototype concept library

Usability

1. learnability: Some participants 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, while 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 comparing of research outcomes. Other participants stated that they need a predefined list and a uniform approach describing

how to use existing codes of additional diagnosis 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 whole used methods. They stated that if they use a code list for each co-morbidity of a condition, they will build an entirely different score over the years. Therefore, transparency in 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 (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 on the basis of 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 in knowing whether there is/will be an active user community for the concept library and the used codes. "*It potentially would make your publication more discoverable. If there's a whole community of users using this"* (data scientist 1).

5) Participants' recommendations to improve repeatable research

Nine participants suggested that the prototype concept library be both UK and globally accessible and practically available to enable researchers around the world to use an online secure platform, which stores codes and other logic, and to encourage researchers to contribute their codes to promote research." *Should be open for the UK"* (data scientist 9). However, one participant recommended that the prototype concept library should be closed at the beginning to ensure it's 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), and in order to know who is using the concept library, data scientist 8 suggested that it should have request sharing, and then 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.

Relating 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).

One participant suggested the use of pre-authorization 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. That approach has many advantages for both the researcher and the publisher, as it means improving the quality of their output. Another participant recommended the creation of a discussion forum in the concept library to facilitate collaboration between 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) Participants' perceptions of their current phenotyping system

The participants mentioned several problems associated with their current system for phenotyping. For example, they have to search for codes from different databases, which utilize different coding systems such as read codes and ICD10 codes, 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 online, they still had to locate the list manually, copy it, and enter those codes into their scripts. Often, they might spend a few days on it and they might miss obscure codes or even use some codes that are irrelevant. *“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). Relating to this, some participants said that they prefer to use code lists, which are referenced and/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 the General Practitioners. They found, for example, 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).

7) Participants' usage and perceptions of existing concept libraries
Not all the participants had previously used some of the existing concept libraries. However, the majority of those who used some of them expressed negative perceptions. For example, several participants stated that the concept libraries they had 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 needed. A further problem for some participants was the inconsistency of data between 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 would like to explore the 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 not to use 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 specialised 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

Statement of the main findings

Development of a concept library that meets users' expectations is extremely useful for repeatable research (e. g., researchers would be able to use the archived code lists to compare studies). This study found that although in principle, everyone felt 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), and this means it has to be very simple (point and click) for some, but have the software and usability to manipulate and design phenotyping algorithms for more advanced users. Also, 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 and this needs 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. In order for the prototype concept library to work, researchers have to engage with it and actually upload their codes there so 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 and barriers to participants sharing their work and reusing the work of others.

The focus group findings demonstrate that facilitators of the participants' sharing of their research methods vary across four categories, namely: 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 [33-35], 2) Benefits for

their research team (e.g., sharing information to promote research within their team) [36,37] , 3) Benefits for their organization (e.g., collaboration between 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) [38]. Relating to this, Crain 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"* [39].

There were several barriers that could inhibit the participants from sharing their research methods, such as expected performance of the shared methods (e.g., 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 (e.g., they were worried about not being referenced by researchers who used their methods). Relating to this, Molloy et al. reported that researchers can be discouraged from sharing their work by fear of not obtaining sufficient credit [41]. Therefore, a safeguard against uncredited use is necessary [42]. In addition, participants mentioned that they are afraid that their methods will be used by other researchers as their own before publication. The results of the study conducted by Xiaolei Huang et al indicated that while most participants are interested in sharing papers related to biodiversity data, more than 60 percent of participants are reluctant to share primary data before publication [43]. 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 many obstacles are resolved, the sharing of data in concept libraries is unlikely to increase significantly.

Several facilitators would encourage the 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 employing such codes. These findings are consistent with those of previous studies. For example, Anneke and Helen reported that researchers are using open research data in order 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 lack of accessibility of other researchers' work, such as code lists, could affect research outcomes, reusing publicly available code lists consumes a lot of time and requires lots of work [38], lack of confidence in online 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 is lacking confidence, 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 (e. g., they do not provide training and/or user documentation, and they are difficult to use) [36-38]. Missing knowledge of the existence of concept libraries and the recognition of how to use them is generally described as an obstacle to data sharing [47]. Because not all researchers use existing concept libraries, obstacles that inhibit researchers from using them need to be addressed when building new concept libraries.

Strengths and limitations

This is the first research, to our knowledge, aimed at identifying the needs of the various users of a concept library. Findings from this study would have a significant impact on improving the efficiency of the 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 in order to improve access to and collaborations with EHRs' routine data, which is part of an all-UK agenda, and the finding of this work will have implications for other countries working to access and share EHRs' routine data.

This study has some limitations that could 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 of the interviewees and focus group participants were recruited because they use the SAIL databank, a national e-health data linkage infrastructure in Wale, so they mostly talked about the Swansea concept library in the SAIL databank. Because the discussion focused on the SAIL databank, its generalization to other concept libraries was limited.

Conclusions

In conclusion, while 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 concepts 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 developing 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 (e.g., ensuring uploading of adequate documentation) for the public good rather than for personal gains.

Acknowledgements

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

Statement on 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. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, et al. Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference. JMIR Med Informatics. 2014;2(2).
2. 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
3. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. 2013;117–21. Available from: https://doi.org/10.1136/amiajnl-2012-001145
4. 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
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 Informatics. 2021;9(8).
6. 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
7. 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
8. 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
9. Kim HH, Kim B, Joo S, Shin SY, 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;21(8):1–11.
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;1(1):53–68.
11. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. J Am Med Informatics Assoc. 2013;20(E2).
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 [Internet]. 2016;71:57–61. Available from: http://dx.doi.org/10.1016/j.artmed.2016.05.005
13. 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
14. 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.
15. 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.
16. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: A new general practice database for research. Inform Prim Care. 2004;12(1):49–50.
17. 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
18. Al Sallakh MA, Vasileiou E, Rodgers SE, Lyons RA, Sheikh A, Davies GA. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. Eur Respir J [Internet]. 2017;49(6). Available from: https://doi.org/10.1183/13993003.00204-2017
19. 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
20. 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
21. 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
22. 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
23. 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
24. 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
25. 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
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 [Internet]. 2018;3(1):10–3. Available from: https://dx.doi.org/10.23889%2Fijpds.v3i1.419
27. 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
28. 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.
29. 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
30. 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
31. 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
32. Braun and Clarke. Successful Qualitative Research: A Practical Guide for Beginners - Virginia Braun, Victoria Clarke - Google Books. 2013;382.
33. 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
34. 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
35. 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
36. 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
37. 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
38. 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
39. 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
40. Stephen J . Ceci Stable. Scientists ' attitudes toward data sharing. Sci Technol Hum Values, Winter - Spring [Internet]. 1988;13(1):45–52. Available from: url: http://www.jstor.com/stable/690052%0A
41. 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
42. 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
43. 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
44. Costello MJ. Motivating Online Publication of Data. Bioscience. 2009;59(5):418–27.
45. Parr CS. Open Sourcing Ecological Data. Bioscience [Internet]. 2007;57(4):309–10. Available from: https://doi.org/10.1641/b570402
46. 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
47. 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
48. Breeze JL, Poline JB, Kennedy DN. Data sharing and publishing in the field of neuroimaging. Gigascience [Internet]. 2012;1(1):2–4. Available from: https://doi.org/10.1186/2047-217X-1-9

Abbreviations
BNF: British National Formulary
CPRD: Clinical Practice Research Data Link
EHRs: Electronic Health Records
ICD: International Classification of Disease
NHS: National Health Service
QOF: Quality and Outcomes Framework
SAIL: Secured Anonymized Information Linkage
SNOMED: Systematized Nomenclature of Medicine
SQL: Structured Query Language
SWOT: Strengths Weaknesses Opportunities Threats
THIN: Health Improvement Network

Supplementary Files

Figures

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 amongst 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).