Clinical code sets and the problem of redundancy in code set repositories

Sigfried Gold [a, b, *], Harold P. Lehmann [b], Lisa M. Schilling [c], Wayne G. Lutters [a]

[a] University of Maryland, College of Information Studies, 4130 Campus Dr., College Park, MD 20742, USA

[b] Johns Hopkins 2024 East Monument St. Suite 1-200. Baltimore, Maryland 21205, USA

[c] University of Colorado Denver - Anschutz Medical Campus, Division of General Internal Medicine,1635 Aurora Ct, Aurora, CO 80045, USA

[*] Corresponding author at: 111 Lee Ave, Apt 106., Takoma Park, MD 20912, USA. *Email address: sigfried@sigfried.org.*

Abstract

Objective: Crafting high-quality code sets is time-consuming and requires a range of clinical, terminological, and informatics expertise. Despite widespread agreement on the importance of reusing code sets, code set repositories suffer from clutter and redundancy, greatly complicating efforts at reuse. When users encounter multiple code sets with the same name or ostensibly representing the same clinical condition, it can be difficult to choose amongst them or determine if any differences among them are due to error or intentional decision.

Methods: This paper offers a view of code set development and reuse based on a field study of researchers and informaticists. The results emerge from an analysis of relevant literature, reflective practice, and the field research data.

Results: Qualitative analysis of our study data, the relevant literature, and our own professional experience led us to three dichotomous concepts that frame an understanding of diverse practices and perspectives surrounding code set development:

1. Permissible values versus analytic code sets;
2. Prescriptive versus descriptive approaches to controlled medical vocabulary use; and
3. Semantic and empirical types of code set development and evaluation practices and the data they rely on.

This three-fold framework opens up the redundancy problem, explaining why multiple code sets may or may not be needed and advancing academic understanding of code set development.

Conclusion: The paper catalogues the methods and practices used and which are appropriate in different contexts. It provides practical aid in managing the code set development process and exposes opportunities for innovation in software to support recommendations made here and in prior literature and to help users navigate thickets of ostensibly redundant code sets not just to choose between them, but to make use of their differences in crafting code sets appropriate to researchers' needs.

Keywords

Clinical code sets; value sets; real-word data; electronic phenotypes; reuse; repository redundancy

1. Introduction

Controlled medical vocabularies (e.g., ICD10, SNOMED, RxNorm, CPT, LOINC) catalogue clinical concepts and relationships between them. A concept is signified by an entry in a medical vocabulary generally consisting of a definition, one or more synonymous labels, and a *code* to identify the concept in representing specific clinical events in electronic health records (EHR), registries, claims databases, and clinical data warehouses. Clinical code *sets* are groupings of these identifiers that facilitate data collection, representation, harmonization, and analysis.[1]

This paper particularly focuses on the use of code sets in the context of observational research using real-world data (RWD) [1]. Despite the use of hierarchical classifications and other data structures to signify concepts at different levels of granularity, code sets are almost always needed when querying clinical data sets since a phenomenon of interest can usually be indicated using a variety of different codes. A study algorithm to determine the relative likelihood of outcome O in patients experiencing condition C depending on their receiving treatment T_1 or T_2 will need to define cohort or phenotype[2] algorithms for identifying patient records indicating O, C, T_1, and T_2 [5]. An essential step in such algorithms is to select patient records containing specified fields whose values are any of the codes in a code set. Though further temporal and conditional logic are often needed beyond the simple presence or absence of matching records in a patient's digital chart, clinical code sets are usually the starting point for phenotype or cohort algorithms.

Crafting high-quality code sets is time-consuming and requires a range of clinical, terminological, and informatics expertise. Scholarly and practical efforts to address challenges in code set management (i.e., helping RWD researchers identify and select the set of codes best fitted to their hypothesis testing and analysis goals) [3–12] have resulted in code set definition and documentation standards [13–17] and in methods and tools for authoring code sets [18–22], for assessing code set semantics and quality [23–29], and for enabling and promoting code set sharing and reuse [30–32]. These papers demonstrate problems of bias and inaccuracy in code sets shared on public repositories and many present specific methods to improve code set development. Williams, et al. [4]—in a paper we used as a seed article for our literature review—performs a comparative review of the code set literature, offering nomenclature, a consolidated articulation of published knowledge on code sets, and a valuable catalog of recommendations for advancing technology for managing code sets.

The current paper offers a view of code set development and reuse based on a field study of researchers and informaticists. We conducted an online survey, semi-structured interviews with a subset of survey participants, and observation where possible of participants working on code sets, finding a diversity in real-world code set development practices and perspectives previously unexplored in the literature.

---

[1] We favor the term "code set", which we treat as more or less synonymous with "value set", "concept set", "code list", and "enumeration", which are also used in some contexts.

[2] Our use of the term electronic phenotype, or just phenotype, follows others in the field of observational research with RWD (e.g., [2–5]) and can be confusing for those not accustomed to this usage. See Section 3.3 for a definition.

While there seems to be universal agreement on the importance of reusing code sets (or phenotype definitions containing code sets), we have recognized through interviews and our own experience that repositories of these objects suffer from clutter and redundancy, greatly complicating efforts at reuse.

Code set repositories tend to contain many code sets with the same name or ostensibly representing the same clinical condition, making it difficult for potential re-users to choose amongst them. When multiple code sets are found, it can be difficult to tell if they are redundant, that is, if any differences among them are due to error or if there are principled reasons to define multiple code sets for certain phenomena.

Qualitative analysis of our study data, the relevant literature, and our own professional experience led us to three dichotomous concepts that frame an understanding of diverse practices and perspectives surrounding code set development. These three dichotomies distinguish:

1. Permissible values versus analytic code sets (which respectively map to primary and secondary uses of clinical data);
2. Prescriptive versus descriptive approaches to controlled medical vocabulary use; and
3. Semantic and empirical types of code set development and evaluation practices and the data they rely on.

We will show how this three-fold framework opens up the redundancy problem, explaining why multiple code sets may or may not be needed (see 3.6.) Our field needs innovative software to help users navigate thickets of ostensibly redundant code sets not just to choose between them, but to make use of their differences in crafting code sets appropriate to researchers' needs.

1.1. Statement of significance

Problem. Despite widespread agreement on the importance of reusing code sets, code set repositories suffer from clutter and redundancy, greatly complicating efforts at reuse.

What is already known. Literature on code set quality and reuse has not recognized the difficulties potential re-users face in leveraging the contributions of prior code set developers.

What this paper adds. This study offers a theoretical framework and practical aid in managing the code set development process by cataloging the methods used and which are appropriate in different contexts. It exposes opportunities for innovation in software support.

2. Methods

As noted, the intent of this research effort is to more deeply understand the diversity of real-world code set development practices, especially mapping the influence of specific contextual factors to those practices. The intended outcomes are both theoretical — developing a more precise, informative set of distinctions between approaches — and practical — providing guidance to informaticists to be deliberate in their decisions, thus enabling more accessible opportunities for both code set and process reuse in the RWD research community.

Our study design has been guided by the scholarly tradition of computer-supported cooperative work (CSCW). It is first predicated on the lived experience of the authors as reflective

practitioners[33] with decades of experience creating and managing code-sets in research contexts. Their initial insight was bolstered or challenged through triangulation among three specific data collection activities: surveys, interviews, and participant observation.

Firstly, a custom, 21-question, web-based survey investigated participants' experiences using clinical code sets in the analysis of RWD. Recruitment focused on professionals with such experience, identifying them through the first author's professional networks. Given the variety and inconsistency in nomenclature for RWD analysis elements and processes, questions were carefully balanced to capture differences in interpretation and use.

Secondly, a sub-set of survey participants were invited for a follow-on semi-structured interview. The purpose was to explore their code set authoring and reuse practices. The contextual nature of the interviews allowed them to demonstrate their tools and processes for developing code sets in person or via screen share.

Thirdly, the first three authors acted as participant observers, embedded in key communities and numerous projects in this space, including OHDSI, PCORNet, Health Data Compass, the Army Pharmacovigilance Center, and the American Medical Informatics Association. While writing this paper, SG and HL worked on the National COVID Cohort Collaborative (N3C), observing and contributing to large-scale code set development and management efforts in a novel context. Their active participation in this wide range of projects has made them careful observers of code set development and curation practices.

The qualitative data collected from the surveys, interviews, and participant observations were content coded in NVivo through a process of analytic induction. Codes and emerging themes were iteratively developed with co-authors. This human subjects research was approved by the University of Maryland IRB (#1405794-8).

In this paper we present the unfolding interpretation of the results of this study as a dialogue among the literature, the reflective practice, and the field research data. The resultant theorizing yields a conceptual framework that is both *descriptive* (making sense and ordering the world as it is) and *prescriptive* (giving structure to practice to inform the world as it ought to be).

3. Results and discussion

3.1. Diversity of code set development contexts

Seventy survey invitations were sent out. Of the 49 responses, 36 were complete enough for analysis, yielding a response rate of 64% and completion rate of 47%. Table 1 shows the diversity of our sample population in terms of relevant demographic and work environment characteristics. Participants hold an array of degrees and work in a variety of disciplines. Most reported being involved in a small number of studies requiring code set development each year, working in teams of between 2 and 10 people, often from multiple organizations. Participants and their fellow team members brought a range of skills and expertise to these projects (Table 1 and Table 2) and they worked on projects involving a range of vocabularies, domains, and data models (Table 3.)

Of the nine most common tools our respondents reported using for code set development listed in Table 3, several are specifically designed for clinical or clinical research applications, providing support for authoring, sharing, and using code sets. The others are general programming and analysis tools with which code sets can be composed, evaluated, and used by linking to database resources containing vocabulary and patient information.

Participant Degrees	#
PhD	17
MS/MA	6
MD	5
MPH	2
BSN	2
RN	1
JD	1
Sector/industry	#
Academic	12
Public	9
Academic professional	8
Pharma	3
Consulting	4
Discipline	#
Informatics	22
Clinical quality measurement	8
Economics	3
Software, epidemiology, ontology	1 each

Studies conducted per year	#
0 to 1	1
1 to 5	2
More than 5	1
Team size (people)	#
1	0
2 to 5	2
5 to 10	9
More than 10	1
Team size (organizations)	#
1	1
2 to 5	2
More than 5	5

Study roles	Any team member	Participant alone	Participant with others	Others	No one
Analyst	35	11	26	10	1
Programmer	35	10	20	15	1
Statistician	33	5	14	19	1
Clinical expert	31	3	5	26	3
Informaticist	30	11	22	7	4
Investigator	30	18	22	10	1
Epidemiologist	25	2	7	17	6
Terminologist	20	6	13	8	13

Table 1. Participant demographics and work contexts; study and/or code set development roles played by participant and other team members

Software/tools used	#
R	26
SQL database	24
OHDSI/ATLAS	17
SAS	13
Python	8
Tableau	7
Epic	6
VSAC	5
i2b2	4
Other	13

Table 2. Software tools, platforms, and repositories used in code set development and sharing; green-shaded items are general programming or analysis tools, blue-shaded items are made particularly for working with medical data or code sets.

Vocabularies used	#	Data models used	#
ICD10CM	29	OMOP	18
CPT	27	PCORNet	9
ICD9CM	26	Local system	9
LOINC	26	Claim forms	5
SNOMED-CT	25	i2b2	4
RxNorm	24	Other	9
HCPCS	22	**Code set domains**	**#**
NDC	21	Conditions	34
OHDSI/OMOP	14	Procedures	30
UMLS	13	Medications	29
MedDRA	8	Lab tests	28
PCORNet	8	Other	9
FDB	5		
Other	19		

Table 3. Vocabularies, vocabulary domains, and data models targeted by participants' code sets.

3.2. Diversity of code set development practices

Code sets and processes for developing them vary in many critical ways. The effectiveness of a given code set development process and the accuracy of the code set it produces depend as much on the thoroughness with which methods are applied as on the selection of those methods.

A particularly important factor shaping code set development practice is whether the code set is being developed for a single project, for use across multiple known projects, or for sharing and reuse in unknown future projects. Literature cited in the introduction [7,8,14,16–18,20,23–28] asserts the importance of reuse in addressing problems with code set quality. We suspected at the outset of this study that reuse was uncommon, as there is considerable evidence [9] that reuse is fraught with difficulties and that repositories accumulate many code sets ostensibly representing the same clinical phenomenon. Our field data (see Table 4) show 30 (83%) of our respondents reuse code sets made by others and 20 (55%) use repositories to find code sets for reuse. Many participants mentioned sharing code sets to public or private repositories — about a third to the Observational Health Data Sciences and Informatics (OHDSI) ATLAS web interface [34], a third to the Value Set Authority Center (VSAC) [31], and several to other repositories, publications, and research networks.

3.3. Permissible values versus analytic code sets

We distinguish two general types of code set based on the way they are used: *permissible value sets*, used for capturing clinical data in patient records, specifying code systems and code system values that can be entered into a particular data element. The items in a permissible value set might be presented to the user as a dropdown list or typeahead field, serving both to prompt the user with the allowable selection of values and prohibit entry of values not included in the set. *Analytic code sets*, on the other hand, are used in the analysis or querying of existing patient records to select those that are indicative of a clinical observation or event of interest where that phenomenon might have been captured using any of a number of codes.[3] There are other use cases for code sets (see Table 5), but the differences between these two contexts (data capture and RWD analysis) will show why the distinction is needed.

[3] In other contexts, such as data harmonization and clinical quality measures, code sets are used in more ambiguous ways that have both permissible and analytic qualities.

Code set-related tasks	#
Create code sets	
Create code sets for local use	28
Create code sets for use by others	27
Vocabulary search	15
Vocabulary navigation	16
Expert advice	16
Optimize code set to more parsimonious expression by replacing codes with intensional rules where possible.	5
Translate code sets across terminologies	21
Add code sets to repositories	24
Use existing code sets	
Use code sets created by others	30
Use code sets from repositories	20
Use code sets from publications	8
Manual or automated comparison with existing code sets	29
Evaluation	
Evaluate code sets (code-by-code or at the set level)	26
Approval for use by subject matter experts or other decision makers.	15
Review by terminology and clinical experts.	18
Empirical valuation	
Examine frequently occurring codes in patients with phenomenon for possible inclusion.	2
Identify false negatives if a reference standard is available.	7
Identify false positives through chart review of matching patients.	12
If codes are semantically appropriate but absent in the intended data, they may be discarded as irrelevant or included for the benefit of future use. (Precalculated term usage counts sufficient.)	2
If prevalence of the target clinical condition is known, a significant discrepancy between prevalence and phenotype/code set result counts may be taken by developers to mean the code set requires further work before release. (Requires patient-level.)	2
Inspect patient and record counts.	15
Review patients matched by code(s) to confirm phenomenon.	11
Sanity check and review the count of patients matched by the whole code set — preferably by executing its containing phenotype algorithm. (Requires access to patient-level detail to generate code set counts.)	13
Sensitivity analysis of changes in results caused by modification of the code set.	4
Test code set on patient data, presumably in context of phenotype algorithm.	13
Other	
If reference standard was used, report how it was made.	2
Report description and justification for validation methods used and any resulting statistics.	3
Informatics, standards, or infrastructure work related to code sets	22

Table 4. Code set-related tasks performed by survey respondents or their team members.

The distinction is not about the digital structure of code sets or their definitions, but about the ways they are used and the practices appropriate to the development and validation of each type. While the distinction is not generally made in the literature or in code set repositories, our findings cannot be understood without drawing it. The HL7 definitions and other discussions of value sets tend to imply permissible as their archetypal use case [23,24,26,27,31,35–38,30,39–43]. While this paper covers both types, analytic are our primary focus [2–4,9,13,32,44–46], and a central claim we make is that analytic code sets necessitate different methods and tools to author, validate, share, and reuse code sets.

Figure 1 shows permissible value sets in context: a clinical data management system includes screens or forms, each of which will include data elements for capturing clinical phenomena like diagnoses, observations, and treatments. Data elements are defined in part by the values they are allowed to take. Specific screens and data elements in EHR, clinical trial, or registry applications may be focused on particular clinical phenomena such as diabetes complications or hypertension medications. A permissible value set then provides a list of subcategories or instances — e.g., cardiomyopathy or retinopathy, etc. for a diabetes complications data element — to populate dropdowns and constrain data element values.

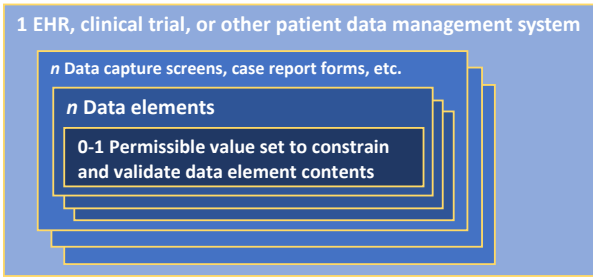

Figure 1. Permissible value sets in context.

In RWD studies, analytic code sets are used in the definition and identification of clinical phenomena of interest, representing study variables such as exposure and comparator cohorts, treatment or exposure criteria, process and outcome, covariates, confounders, etc.[3] The algorithmic components that identify specific clinical phenomena in the data may be called electronic phenotypes, phenotype algorithms, cohort definitions, or just variables; this paper mostly refers to them as "phenotypes." Phenotype algorithms may use various types of data (narrative notes, images, EKG or other device output, etc.), but insofar as terminology codes are used in the algorithm, a phenotype will include one or more code sets as diagrammed in Figure 2 and may use temporal and conditional logic in performing set operations on the groups of patient records matched by different code sets. However, phenotypes can also be as simple as a single code set, the algorithm consisting of nothing but the selection of patient records containing one of the codes in that set.[4]

[4] See, e.g., "Finding Existing Phenotype Definitions" in the phenotyping chapter of the online textbook, Rethinking Clinical Trials [2] which lists code set repositories alongside repositories of more complex algorithms as sources of reusable electronic phenotypes.

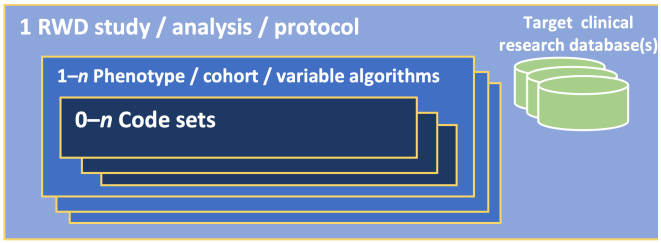

Figure 2. Analytic code sets in context

Permissible value sets are generally developed for specific clinical data systems, often for a single institution. Analytic code sets are usually designed to reflect nuances of a particular research question, such as a need for sensitivity or specificity or the need for study-specific exclusion criteria.

3.4. Prescriptive and descriptive perspectives on code sets

Distilling and analyzing the catalogue of code set development practices in Table 4 — how code sets are made, used, evaluated, and reused — led us to re-evaluate the literature around code sets, observing that it falls into three research focus areas, which often do not seem to be in dialog with each other:

1. On value sets for permissible values or clinical quality measures (CQM) [23–27,30,31,43,35–37,39–42,47];[5]
2. On phenotypes (cohort selection) for RWD research [2,3,13,45,46]; and
3. On (analytic) code sets for use in phenotyping applications [4,9,17,19,32].

The computable phenotype literature sometimes conflates phenotypes and analytic code sets. Where it does discuss reuse, the focus is on phenotype rather than code set reuse. While some in this group do not believe that sharing code sets separately from phenotypes is worthwhile, we have seen that reuse does occur and there is demand for code sets that can be used across phenotypes. The value set and code set literatures both tend to focus on code set repositories and reuse.

The value set literature is not concerned with adapting code sets to specific clinical databases and looks to expert review or published sources of authority for code set validation. The phenotype literature, on the other hand, evaluates code set correctness primarily through empirical analysis with clinical data. The code set literature falls somewhere in between.[6]

Whether one considers expert authority or empirical analysis as the primary means of code set evaluation can reflect almost ideological beliefs about the nature of code sets and medical

[5] Although value sets are not used as permissible value constraints in CQMs, the value set literature does not treat them any differently and does not address issues around aligning them with patient data — perhaps because CQM use by health care organizations is often required by payers, so alignment of value set and patient data is the responsibility of provider organizations, not value set developers.

[6] The current paper shares more in common with this third group than the others, but it draws on all three. Those who identify, explicitly or implicitly, with one of these groups will benefit from being also informed by the others.

vocabulary use generally. Outside the domain of medicine and controlled medical vocabularies, in lexicographical or grammatical terms, the dichotomy between *prescriptivist* views and *descriptivist* views is well-known. For prescriptivists, dictionary entries and grammatical rules define how words and language *should be* used; proper language should conform with such rules and definitions. For descriptivists, dictionaries and grammars are attempts to capture a snapshot of how words and language *are* used in a given time and milieu. Non-conformant usage patterns indicate that the rules are lacking, not that the usage is wrong.

While terminological prescriptivism in natural language is generally considered unscientific and pedantic [48], the imposition of prescriptive terminology is, of course, the foundational purpose of standardized medical vocabularies and arguably the foundational practice of medical informatics generally [49–52].

Exemplifying a prescriptivist orientation, Winnenburg, et al. 2013 suggests that a code set should be anchored in a single concept, consisting of that concept and its descendants. That view is rejected by the descriptivist perspective held by many RWD researchers, expressed in rather extreme terms by one survey respondent:

> Code sets are always context specific. There is no such thing as diabetes in an RWD data source, there might be 50 definitions of diabetes and you have to pick the one that matches your question, data, and methods…We may spend months developing a code set for a specific question, iterating on different algorithms until the investigator is satisfied that the definition matches the needs of the study. (P04)

The following section distinguishes semantic from empirical techniques in code set development and validation. We have observed in a minority of our participants and in some of the literature a bias towards either semantic or empirical techniques that seems to override consideration of context and to reflect an implicit commitment to prescriptivist or descriptivist perspectives on controlled vocabulary use. That is, a person holding one of these views can find it difficult to see merit in the alternative. Overall, prescriptivist views should be appropriate for permissible value sets and data capture contexts; and descriptivist for analytic code sets and RWD research.

3.5. Semantic versus empirical methods and resources

Williams 2017 alludes to a central tension in the choice of methods for code set validation [4]. On the one hand, a rigorous validation would be to compare patient selection results against a reference ("gold") standard created through medical record abstraction (MRA.) On the other, creating such a standard can be prohibitively time-consuming and require data that may be challenging or impossible to obtain. In our data, evaluation fell into two major categories, which we label overall as semantic and empirical.

3.5.1. Semantic methods and evaluation by authority

We asked participants, "How do you verify that you have selected the best codes for representing a clinical concept in your analyses?" and received a range of answers. For many, confidence in their code selection came by reusing existing code sets from "previous[ly] published results"

(P01), "examin[ing] the literature for validation studies" (P41), or "validated codesets when possible" (P04).[7]

Other participants described an evaluation or validation process based on review by terminologists and clinicians (P19), clinical experts (P26), or "our coding panel, a group of experts that give us advice and feedback" (P21).

3.5.2. Patient data and empirical evaluation

Many participants consider these semantic evaluation methods — reuse of existing concept sets and expert review — sufficient; others, however, hold that a code set for analytic use cannot be trusted without evaluating it or the phenotype or algorithm containing it through some form empirical review of clinical data:

> Chart Review. Some internal checking of codes against expected lab results, vital measurements, patient histories, etc. Ex. Diabetes codes should associate with histories of certain blood glucose measurements or A1C. (P28)

> First conduct discussion with clinical experts; Second, evaluate coverage of clinical concept in a data set; Third, perform random chart review to help detect if presence of code indicates disease. (P16)

> Lexical search, semantic exploration (navigate OHDSI vocab), empirical assessment thru characterization, and clinical expert review. (P05)

Every one of our 36 survey respondents reported that their *studies* use patient data. Nevertheless, of the 32 who answered the open-ended validation question, only nine indicated using patient data during code set development and validation. Choice of evaluation methods can be guided by clinical nuances of the research question or how the code set will be used. According to one survey respondent:

> Depends on the purpose and whether we are aiming for sensitivity or specificity. It may be chart review, or comparison with other value sets. (P34)

To unpack that statement a bit, a highly *sensitive* code set might be appropriate for instance for selecting patients to be screened for some condition where the goal is to capture as many patients as possible. When a sensitive code set is needed, comparison with existing code sets could help to make sure that appropriate codes are not missed. A highly *specific* code set may be suitable when recruiting patients for a clinical trial or when constructing the main cohort for an observational study, where false positives are costly. In this case, chart review of a sample of identified patients and code set/phenotype modifications will help.

---

[7] Code sets are available from code set repositories such as VSAC, ClinicalCodes [32], or the OHDSI/ATLAS or N3C concept set editors [34,57,58]); published papers that follow RECORD and other data-based observational study reporting guidelines [13,15]; previous projects available to the code set developer; and groupers such as Clinical Classifications Software Refined (CCSR) [59,60].

Evaluating the accuracy of a sensitive code set requires a thorough semantic exploration to identify all codes that could indicate the condition of interest,[8] while evaluating the accuracy of a specific code set should involve empirical examination of matching patient records to prevent false positives.

Table 5 lists general contexts in which code sets are used and relates them to the three literature categories listed above, to code set repositories that support them, and to the three conceptual dichotomies described in Sections 3.3, 3.4, 3.5: permissible/analytic, prescriptive/descriptive, and semantic/empirical.

[8] Software tools are available particularly to help code set developers discover codes related to the ones they start with: Term Sets [17] and PHenotype Observed Entity Baseline Endorsements (PHOEBE) [53]. PHOEBE functionality has recently been added to OHDSI's ATLAS concept set editor.

Context	Literature group	Reuse repositories	Permissible / analytic	Prescriptive / descriptive	Semantic / empirical evaluation
Value sets for data capture	Value set	VSAC	Permissible	Prescriptive	Semantic
Other value sets for terminology services (FHIR, CTS2, etc.)	Value set	VSAC	Permissible	Prescriptive	Semantic
Clinical quality measures	Value set	VSAC	Both	Both	Mostly semantic
Single study, single database	Phenotype		Analytic	Descriptive	Both — but empirical is vital and possible
Network study or multiple related studies	Phenotype	ATLAS, N3C	Analytic	Descriptive	Both
For analytic reuse but not for a specific study, database, or question	Code set	N3C	Analytic	Descriptive	Mostly semantic

Table 5. Contexts for code set development

3.6. A taxonomy of reasons for code sets to differ

While code set repositories make it possible to share and reuse code sets, clutter and redundancy can present serious challenges. For instance, a search for COPD (chronic obstructive pulmonary disease) on ATLAS (https://atlas-demo.ohdsi.org/#/conceptsets) gives 56 results. While many of these are usefully distinguished by their titles (e.g., Stage III-IV COPD or Concomitant COPD), many are not. In this section we break down the reasons that code sets for (ostensibly) the same clinical concept may differ into three categories: semantic, empirical, and due to error. These are summarized in Table 6.

3.6.1. Semantic reasons for code sets to differ

Two code sets may have the same name and refer to ostensibly the same condition or event but, on closer inspection, may differ in their meaning or how they are meant to be used.

Clinical meaning or nuance. Researchers may differ in their understanding of a clinical concept, or, for instance, a diabetes code set for a cardiology study may require a different set of codes than an endocrinology study.

Study requirements. Different code sets may share the same clinical meaning, but one study may need a more specific code set, another a more sensitive one.

Vocabulary issues. Code sets for the same phenomenon will, of course, differ if they use different vocabularies (e.g., ICD10CM or SNOMED-CT for clinical findings; NDC, RxNorm, or ATC for drugs; CPT, HCPCS, or ICD10PCS for procedures.) There may be reasons to translate codes across vocabularies.[9] Different strategies may be applied when using vocabularies that either lack granularity to express the concept of interest or use post-coordination to express it.

Algorithmic context. Different strategies for identifying a clinical phenomenon can lead to differences in code set composition. (This not strictly a semantic issue but a fact of code set use.)

- A code set designed to target a diagnostic condition might use evidence from other domains of clinical data (e.g., drugs, procedures, lab tests.)[10]
- A code set may be designed in the knowledge that it will produce false positives or negatives if these will be corrected by logic or other code sets at the phenotype.

3.6.2. Empirical reasons for code sets to differ

Differences in the datasets being analyzed may affect the codes used to represent some conditions.

Population characteristics. E.g., codes may differ for studies of children less than 10 years of age versus geriatric populations; a study of neuropathy in orthopedic surgery patients would not need to include codes for diabetic neuropathy in the pediatric population.

---

[9] E.g., CDMs like OMOP may require translation or harmonization of codes in patient records to agreed-upon vocabularies.

[10] It is recognized in the literature that phenotypes benefit from the use of multi-modal, multi-domain data[61].

Regional, institutional, or clinical specialization coding practices. A single meaning may be expressed using different codes in different places.[11]

Institutional workflow at data source. Certain conditions or observations may not be captured in some clinical settings requiring recourse to indirect ways of identifying them in EHRs.

Arbitrary inclusion thresholds. Some codes may give rise to false positives when included and false negatives when left out. If researchers are not able to resolve this kind of problem at the phenotype algorithm level, they will need to make judgment calls depending on whether they think the false negatives or false positives caused by a given code's presence or absence will more adversely affect the study's results. Differences in judgment do not mean one decision is right and another wrong, but, unless the judgment call is justified with specific reasoning, code sets for a given phenomenon can differ without giving potential re-users any basis for choosing between them.

3.6.3. Errors

Crafting accurate code sets is hard and mistakes are not uncommon. Discrepancies between code sets provide an opportunity to discover mistakes that might otherwise be overlooked.

Codes mistakenly left out or mistakenly included. When code sets are missing codes they should include, they can cause false negatives in patient or event selection; codes included in error can introduce false positives in selection. Without a reference (or gold) standard to test results (of a code set or its containing cohort algorithm) against a sample of records already reliably classified as exhibiting or not exhibiting the clinical phenomenon of interest, false positives and negatives in selection results may entirely escape detection.

Codes included or not based on faulty or idiosyncratic reference standards. Reference standards themselves can suffer from error. Decisions by a chart reviewer on which patients match a phenotype or cohort definition can be affected by differences in understanding that are not quite matters of clinical judgment or study needs but differences in chart reading practice, differences in the chart reviewers' interpretation of study needs, or chart reviewer error. But if the error or discrepancy affects the standard, a code set or its containing phenotype may show perfect sensitivity (low false negative rate) and specificity (low false positive rate) while differing from a code set based on another gold standard.

Errors can lead to bias in results, whose magnitude and direction are not predictable, but legitimate differences in code sets can be recognized if their reasons are known. Code set analysis and authoring software can be better designed to help re-users understand these differences, giving them a basis for deciding between existing code sets or selecting the elements from each most appropriate for their own use case.

[11] This possibility was mentioned by our participants and seems to be a relatively common belief, but we have not encountered specific examples.

<table>
<tr><th>VALID SEMANTIC</th><th>VALID EMPIRICAL</th><th>ERRONEOUS AND ARBITRARY</th></tr>
<tr><td>Clinical meaning or nuance
Study requirements
Terminologies and cross-terminology mappings
Use of vocabularies lacking granularity for clinical concepts or requiring post-coordination</td><td>Population characteristics
Regional, institutional, or clinical specialization coding practices
Institutional workflow</td><td>Codes mistakenly left out
Codes mistakenly included
Codes included or not based on faulty or idiosyncratic reference standards</td></tr>
<tr><td colspan="2">Algorithmic context (use of other domains or correcting for false positives or negatives at the phenotype level)</td><td></td></tr>
<tr><td></td><td colspan="2">Arbitrary inclusion thresholds</td></tr>
</table>

Table 1. Reasons for code sets with same name to differ in definition and composition.

3.7. Conclusion: Leveraging and mitigating redundancy in code set repositories

Code set reuse is frequently championed as a response to persistent concerns about code set quality: not only should researchers make use of expertly designed code sets, code set repositories should facilitate incremental refinement; over time the quality of a shared code set should improve as more researchers put it to use, evaluate its accuracy, and contribute their changes back to the repository.

> [Reusable code sets] would be helpful [so that] I don't have to do this on my own every time…[B]ecause it has been created by a collaborative team that's known for creating value sets, I would know that, "Oh, this has been extracted or they got it from a paper that has been vetted and validated and you know it's a legit paper." I would use that. (P09, interview)

Rather than incremental improvement of existing code sets or indications of a code set's having been vetted and validated, what we see in repositories is proliferation and clutter: new code sets that may or may not have been vetted in any way and junk concept sets, created for some reason but never finished. We have found general agreement in our data that the presence of many alternative code sets for a given condition often leads code set developers to ignore all of them and start from scratch, as there is generally no easy way to tell which will be more appropriate for the researcher's needs. And if they share their code set back to the repository (as they must on analysis platforms like ATLAS or N3C), they further compound the problem, especially if they neglect to document the new code set's intention and provenance.

There is a tension regarding how many code sets should exist for a given clinical condition. On the one hand, the principle of reproducibility of research and fungibility of research results—whether results from different studies may be pooled—argues for re-use of code sets. On the other hand, tight coherence with the research question — "fitness for use" — argues for customizing a unique code set to fit the research intent. Given this tension, it is no surprise that respondents expressed a variety of beliefs on each side of this dialectic.

If, as a field, we hope to increase reuse and refinement to decrease redundant code set creation, we must be able to understand when an additional code set for a target condition may be needed or not. The taxonomy in Section 3.6 may help in reconciling differences when multiple code sets are being reviewed or considered for reuse: if the analyst can identify a valid reason for a difference, this may give them insight to inform choices for their own use case or may help them determine where errors lie, increasing or decreasing their confidence in specific code sets or codes.

When a new researcher creates their own code set from scratch rather than leveraging the work of those who have tread the same ground, however, this should not be seen as laziness or as a problem to be addressed by exhortations to reuse existing code sets. Rather, the fault should be ascribed to the resources available to them: they should be given software and metadata to make the review and comparison of existing code set at least feasible, if not easier than creation from scratch. Practical application of the taxonomy and other ideas presented in this paper will require

new software designed to implement these ideas and better guide code set developers through the process. Toward that end, we offer the following recommendations.

3.7.1. Advanced, automated comparison tools

In our professional experience we have seen instances where trust in what were considered authoritative code sets broke down when comparing them to other code sets. One participant, P16, performed an automated comparison of many alternative code sets for depression, using the differences and similarities to create a trustable code set without having to trust any of the input code sets individually.

Comparison functionality should be a central feature of code set repositories and authoring platforms, allowing users to take advantage of existing code sets rather than burdening them with having to manually sift through ostensibly redundant code sets. In the last couple years, tools for comparing code sets have begun to appear in ATLAS, the N3C Concept Set Browser, and TermHub [62]. TermHub explicitly nudges the user to compare related code sets and highlights the selected code sets' similarities and differences throughout authoring and review.

3.7.2. Detailed metadata collection and use

Existing tools vary in their collection of metadata through the authoring process, but however much metadata they collect, it is at the code set level; it could be enormously helpful to collect metadata to capture code set developers' reasoning for including or rejecting specific codes.[12] A combination of automated process data capture and timely, minimally obtrusive user prompts could provide code-level metadata that could be displayed as future code set authors consider whether to include a code or not. An automated capture process could, for instance, record the source of included codes: if found in an existing code set, record a reference to that code set; if found through vocabulary text search and/or navigation of vocabulary hierarchy, record the steps leading to the included code. User prompts could try, for instance, to capture whether patient counts or any kind of chart review or gold standard had been used in decisions to include or reject specific codes.

3.7.3. Expert curation

Terminology experts on the N3C infrastructure staff have developed "N3C Recommended" code sets for commonly studied topics (conditions, medications, medication classes, measurements, procedures.) TermHub was specifically designed to facilitate that endeavor.

The requirements and recommendations in prior literature have not been sufficient to guide the design of software that could make effective leveraging of shared code sets a reality. However, the conceptual framework, real-world experience, and deep, detailed account of the challenges to reuse presented here make up that deficit and provide a high-level requirements roadmap for improved code-set creation tools.

---

[12] FHIR and N3C accommodate relatively extensive set of metadata fields; VSAC somewhat less; and OHDSI/ATLAS hardly any at all. N3C, at SG's suggestion, does request reasoning when adding codes to a code set, but this feature has not yet been developed to the point of being useful — nothing is currently done with users' input.

CRediT authorship contribution statement

Sigfried Gold: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization; **Harold P. Lehmann**: Writing - Review & Editing, Supervision; **Lisa M. Schilling**: Writing - Review & Editing, Supervision; **Wayne G. Lutters**: Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration

Acknowledgments

Richard Williams, Jessica Ancker, Christopher Chute, Davera Gabriel, Harold Solbrig, Jeff Brown, Laura Wiley, Luke Rasmussen, Rachel Richesson, Allen Flynn, Meredith Zozus, David Gotz, Erica Voss, Chrstian Reich, Kristin Kosta, Shelley Rusincovitch, Niklas Elmqvist, Leilani Battle, Joel Chan, Amanda Lazar.

Funding: Sigfried Gold's contribution to this research was supported in part by NSF award DGE-1632976.

Bibliography

[1] J. Corrigan-Curay, L. Sacks, J. Woodcock, Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness, JAMA 320 (2018) 867–868. https://doi.org/10.1001/jama.2018.10136.

[2] J. Pathak, J. Wang, S. Kashyap, M. Basford, R. Li, D.R. Masys, C.G. Chute, Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience, Journal of the American Medical Informatics Association 18 (2011) 376–386. https://doi.org/10.1136/amiajnl-2010-000061.

[3] G. Hripcsak, D.J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Inform. Assoc. 20 (2013) 117–121. https://doi.org/10.1136/amiajnl-2012-001145.

[4] R. Williams, E. Kontopantelis, I. Buchan, N. Peek, Clinical code set engineering for reusing EHR data for research: A review, Journal of Biomedical Informatics 70 (2017) 1–13. https://doi.org/10.1016/j.jbi.2017.04.010.

[5] R.L. Richesson, L.K. Wiley, S. Gold, L. Rasmussen, Luke V., Electronic Health Records-Based Phenotyping, Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials (2020). https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/electronic-health-records-based-phenotyping-introduction/ (accessed July 14, 2020).

[6] S. Davé, I. Petersen, Creating medical and drug code lists to identify cases in primary care databases: CREATING MEDICAL AND DRUG CODE LISTS USING STATA, Pharmacoepidem. Drug Safe. 18 (2009) 704–707. https://doi.org/10.1002/pds.1770.

[7] R.L. Richesson, S.A. Rusincovitch, D. Wixted, B.C. Batch, M.N. Feinglos, M.L. Miranda, W.E. Hammond, R.M. Califf, S.E. Spratt, A comparison of phenotype definitions for

diabetes mellitus, J. Am. Med. Inform. Assoc. 20 (2013) e319-26. https://doi.org/10.1136/amiajnl-2013-001952.

[8] S. Lanes, J.S. Brown, K. Haynes, M.F. Pollack, A.M. Walker, Identifying health outcomes in healthcare databases, Pharmacoepidemiol Drug Saf 24 (2015) 1009–1016. https://doi.org/10.1002/pds.3856.

[9] S. Gold, A. Batch, R. McClure, G. Jiang, H. Kharrazi, R. Saripalle, V. Huser, C. Weng, N. Roderer, A. Szarfman, N. Elmqvist, D. Gotz, Clinical Concept Value Sets and Interoperability in Health Data Analytics, AMIA Annu Symp Proc 2018 (2018) 480–489. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371254/ (accessed March 11, 2019).

[10] S. Gold, H. Lehmann, L. Schilling, W. Lutters, Practices, norms, and aspirations regarding the construction, validation, and reuse of code sets in the analysis of real-world data, Submitted to the Journal of Biomedical Informatics (2021) 35. https://doi.org/10.1101/2021.10.14.21264917.

[11] G. Hripcsak, D.J. Albers, High-fidelity phenotyping: richness and freedom from bias, Journal of the American Medical Informatics Association 25 (2018) 289–294. https://doi.org/10.1093/jamia/ocx110.

[12] A. Ostropolets, L. Zhang, G. Hripcsak, A scoping review of clinical decision support tools that generate new knowledge to support decision making in real time, Journal of the American Medical Informatics Association 27 (2020) 1968–1976. https://doi.org/10.1093/jamia/ocaa200.

[13] A. Callahan, N.H. Shah, J.H. Chen, Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data, Annals of Internal Medicine 172 (2020) S79–S84. https://doi.org/10.7326/M19-0873.

[14] Vocabulary Work Group, HL7 Specification: Characteristics of a Formal Value Set Definition, Release 1, HL7 ANSI, 2019. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=437 (accessed March 8, 2018).

[15] E.I. Benchimol, L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, H.T. Sørensen, E. von Elm, S.M. Langan, RECORD Working Committee, The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement, PLOS Medicine 12 (2015) e1001885. https://doi.org/10.1371/journal.pmed.1001885.

[16] B.S. Alper, A. Flynn, B.E. Bray, M.L. Conte, C. Eldredge, S. Gold, R.A. Greenes, P. Haug, K. Jacoby, G. Koru, J. McClay, M.L. Sainvil, D. Sottara, M. Tuttle, S. Visweswaran, R.A. Yurk, Categorizing metadata to help mobilize computable biomedical knowledge, Learning Health Systems 6 (2022) e10271. https://doi.org/10.1002/lrh2.10271.

[17] R. Williams, B. Brown, E. Kontopantelis, T. van Staa, N. Peek, Term sets: A transparent and reproducible representation of clinical code sets, PLoS ONE 14 (2019) e0212291. https://doi.org/10.1371/journal.pone.0212291.

[18] K.J. Peterson, G. Jiang, S.M. Brue, F. Shen, H. Liu, Mining Hierarchies and Similarity Clusters from Value Set Repositories, AMIA Annu Symp Proc 2017 (2017) 1372–1381.

[19] J. Watson, B.D. Nicholson, W. Hamilton, S. Price, Identifying clinical features in primary care electronic health record studies: methods for codelist development, BMJ Open 7 (2017) e019637. https://doi.org/10.1136/bmjopen-2017-019637.

[20] R. Williams, GetSet, (2019). https://getset.ga/ (accessed September 13, 2021).

[21] L. Zhang, Y. Zhang, T. Cai, Y. Ahuja, Z. He, Y.-L. Ho, A. Beam, K. Cho, R. Carroll, J. Denny, I. Kohane, K. Liao, T. Cai, Automated grouping of medical codes via multiview banded spectral clustering, Journal of Biomedical Informatics 100 (2019) 103322. https://doi.org/10.1016/j.jbi.2019.103322.

[22] A. Ostropolets, G. Hripcsak, C. Knoll, P. Ryan, PHOEBE 2.0: selecting the right concept sets for the right patients using lexical, semantic, and data-driven recommendations, (n.d.).

[23] R. Winnenburg, O. Bodenreider, Issues in creating and maintaining value sets for clinical quality measures, AMIA Annu. Symp. Proc. 2012 (2012) 988–996. https://www.ncbi.nlm.nih.gov/pubmed/23304374.

[24] R. Winnenburg, O. Bodenreider, Metrics for assessing the quality of value sets in clinical quality measures, AMIA Annu. Symp. Proc. 2013 (2013) 1497–1505. https://www.ncbi.nlm.nih.gov/pubmed/24551422.

[25] R. Winnenburg, L. Rodriguez, F.M. Callaghan, A. Sorbello, A. Szarfman, O. Bodenreider, Aligning Pharmacologic Classes Between MeSH and ATC, in: 2013. https://mor.nlm.nih.gov/pubs/pdf/2013-vdos-rw.pdf.

[26] N.J. Bahr, S.D. Nelson, R. Winnenburg, O. Bodenreider, Eliciting the Intension of Drug Value Sets – Principles and Quality Assurance Applications, Stud Health Technol Inform 245 (2017) 843–847. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5881398/ (accessed May 14, 2020).

[27] R.A. Cholan, N.G. Weiskopf, D.L. Rhoton, N.V. Colin, R.L. Ross, M.N. Marzullo, B. Sachdeva, D.A. Dorr, Specifications of Clinical Quality Measures and Value Set Vocabularies Shift Over Time: A Study of Change through Implementation Differences, AMIA Annu Symp Proc 2017 (2018) 575–584. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977609/ (accessed May 13, 2020).

[28] D. Margawati, UNDERSTANDING THE VARIABILITY IN VALUE SETS: THE ROLE OF STEWARD, Johns Hopkins University, 2019.

[29] J.N. Swerdel, G. Hripcsak, P.B. Ryan, PheValuator: Development and evaluation of a phenotype algorithm evaluator, Journal of Biomedical Informatics 97 (2019) 103258. https://doi.org/10.1016/j.jbi.2019.103258.

[30] O. Bodenreider, D. Nguyen, P. Chiang, P. Chuang, M. Madden, R. Winnenburg, R. McClure, S. Emrick, I. D'Souza, The NLM Value Set Authority Center, Stud Health Technol Inform 192 (2013) 1224. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300102/ (accessed March 9, 2019).

[31] E. Khatipov, M. Madden, P. Chiang, P. Chuang, D.M. Nguyen, I. D'Souza, R. Winnenburg, O. Bodenreider, J. Skapik, R.C. McClure, S. Emrick, Creating, Maintaining and Publishing Value Sets in the VSAC, in: AMIA, 2014.

[32] D.A. Springate, E. Kontopantelis, D.M. Ashcroft, I. Olier, R. Parisi, E. Chamapiwa, D. Reeves, ClinicalCodes: An Online Clinical Codes Repository to Improve the Validity and Reproducibility of Research Using Electronic Medical Records, PLoS ONE 9 (2014) e99825. https://doi.org/10.1371/journal.pone.0099825.

[33] D.A. Schön, The reflective practitioner: how professionals think in action, Basic Books, New York, 1983. http://www.gbv.de/dms/bowker/toc/9780465068746.pdf (accessed August 30, 2023).

[34] OHDSI/ATLAS, (2020). https://github.com/OHDSI/Atlas/wiki (accessed May 5, 2020).

[35] K.J. Peterson, G. Jiang, S.M. Brue, F. Shen, H. Liu, Mining Hierarchies and Similarity Clusters from Value Set Repositories, AMIA Annu Symp Proc 2017 (2017) 1372–1381.

[36] J. Pathak, G. Jiang, S.O. Dwarkanath, J.D. Buntrock, C.G. Chute, C. Chute, LexValueSets: an approach for context-driven value sets extraction, AMIA Annu Symp Proc (2008) 556–560.

[37] H. Solbrig, ISO 11179 CTS2 and Value Set Binding, (n.d.). http://dbooth.org/2015/solbrig/FHIR_RDF_Solbrig.pdf (accessed March 28, 2020).

[38] G. Jiang, H.R. Solbrig, C.G. Chute, Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups, J. Am. Med. Inform. Assoc. 19 (2012) e129-36. https://doi.org/10.1136/amiajnl-2011-000739.

[39] J. Wu, J.T. Finnell, D.J. Vreeman, Evaluating Congruence Between Laboratory LOINC Value Sets for Quality Measures, Public Health Reporting, and Mapping Common Tests, AMIA Annu Symp Proc 2013 (2013) 1525–1532. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900163/ (accessed May 14, 2020).

[40] HL7 Standards Product Brief - HL7 Specification: Characteristics of a Formal Value Set Definition, Release 1, (2017). http://www.hl7.org/implement/standards/product_brief.cfm?product_id=437 (accessed March 8, 2018).

[41] C. Caragea, V. Honavar, P. Boncz, P. Boncz, P.-Å. Larson, S.W. Dietrich, G. Navarro, B. Thuraisingham, Y. Luo, O. Wolfson, S.M. Beitzel, E.C. Jensen, O. Frieder, C.S. Jensen, N. Tradišauskas, E.V. Munson, A. Wun, K. Goda, S. E. Fienberg, J. Jin, G. Liu, N. Craswell, T.B. Pedersen, C. Pautasso, M.M. Moro, S. Manegold, S. Manegold, B. Carminati, M. Blanton, S. Bouchenak, N. de Palma, W. Tang, C. Quix, W. Tang, M.A. Jeusfeld, R.K. Pon, D.J. Buttler, M.A. Jeusfeld, W. Meng, P. Zezula, M. Batko, V. Dohnal, J. Domingo-Ferrer, J. Domingo-Ferrer, D. Barbosa, I. Manolescu, J. Xu Yu, J. Domingo-Ferrer, E. Cecchet, V. Quéma, X. Yan, O. Wolfson, G. Santucci, D. Zeinalipour-Yazti, P.K. Chrysanthis, C. Quix, A. Deshpande, C. Guestrin, S. Madden, C.K.-S. Leung, R.H. Güting, R.H. Güting, A. Gupta, T.B. Pedersen, H. Tao Shen, G. Weikum, B. Thuraisingham, G. Weikum, R. Jain, J.X. Yu, P. Ciaccia, K.S. Candan, M.L. Sapino, J.X. Yu, R. Jain, C. Meghini, F. Sebastiani, U. Straccia, F. Nack, V.S. Subrahmanian, M.V. Martinez, Dr. Reforgiato, J.X. Yu, T. Westerveld, M. Sebillo, G. Vitiello, M. De Marsico, K. Voruganti, C. Parent, S. Spaccapietra, C. Vangenot, E. Zimányi, P. Roy, S. Sudarshan, E. Puppo, P. Kröger, M. Renz, H. Schuldt, S. Kolahi, A. Unwin, W. Cellary, Metadata Registry, ISO/IEC 11179, in: L. Liu, M.T. Özsu (Eds.), Encyclopedia of Database Systems, Springer US, Boston, MA, 2009: pp. 1724–1727. https://doi.org/10.1007/978-0-387-39940-9_907.

[42] National Cancer Institute, caDSR and ISO 11179, Cancer Data Standards Registry and Repository (2016). https://wiki.nci.nih.gov/display/caDSR/caDSR+and+ISO+11179 (accessed May 26, 2020).

[43] FHIR Vocabulary Work Group, ValueSet - FHIR v4.0.1, HL7 FHIR Release 4 (2019). https://www.hl7.org/fhir/valueset.html (accessed December 8, 2020).

[44] R. Williams, B. Brown, E. Kontopantelis, T. van Staa, N. Peek, Term sets: A transparent and reproducible representation of clinical code sets, PloS One 14 (2019) e0212291. https://doi.org/10.1371/journal.pone.0212291.

[45] K.M. Newton, P.L. Peissig, A.N. Kho, S.J. Bielinski, R.L. Berg, V. Choudhary, M. Basford, C.G. Chute, I.J. Kullo, R. Li, J.A. Pacheco, L.V. Rasmussen, L. Spangler, J.C. Denny, Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network, J Am Med Inform Assoc 20 (2013) e147–e154. https://doi.org/10.1136/amiajnl-2012-000896.

[46] X. Jing, M. Emerson, D. Masters, M. Brooks, J. Buskirk, N. Abukamail, C. Liu, J.J. Cimino, J. Shubrook, S. De Lacalle, Y. Zhou, V.L. Patel, A visual interactive analytic tool for filtering and summarizing large health data sets coded with hierarchical terminologies (VIADS), BMC Medical Informatics and Decision Making 19 (2019) 31. https://doi.org/10.1186/s12911-019-0750-y.

[47] G. Jiang, H.R. Solbrig, C.G. Chute, Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups, J. Am. Med. Inform. Assoc. 19 (2012) e129-36. https://doi.org/10.1136/amiajnl-2011-000739.

[48] A Word on "Descriptive" and "Prescriptive" Defining, Https://Www.Merriam-Webster.Com/ (n.d.). https://www.merriam-webster.com/words-at-play/descriptive-vs-prescriptive-defining-lexicography (accessed August 21, 2021).

[49] C.G. Chute, The Copernican era of healthcare terminology: a re-centering of health information systems, Proc AMIA Symp (1998) 68–73. https://www.ncbi.nlm.nih.gov/pubmed/9929184.

[50] H.D.L. Rosenberg HM, History of the statistical classification of diseases and causes of death, National Center for Health Statistics, Hyattsville, MD, 2011. https://www.cdc.gov/nchs/data/misc/classification_diseases2011.pdf.

[51] M. Berg, G. Bowker, The multiple bodies of the medical record, Sociol. Q. (1997). http://onlinelibrary.wiley.com/doi/10.1111/j.1533-8525.1997.tb00490.x/full.

[52] G.C. Bowker, S.L. Star, Sorting Things Out, (2000). https://mitpress.mit.edu/books/sorting-things-out (accessed September 26, 2017).

[53] A. Ostropolets, P. Ryan, G. Hripcsak, Phenotyping in distributed data networks: selecting the right codes for the right patients, AMIA Annu Symp Proc 2022 (2022) 826–835.

[54] K.W. Fung, J. Xu, S. Gold, The Use of Inter-terminology Maps for the Creation and Maintenance of Value Sets, AMIA Annu Symp Proc 2019 (2019) 438–447.

[55] N. Matentzoglu, J.P. Balhoff, S.M. Bello, C. Bizon, M. Brush, T.J. Callahan, C.G. Chute, W.D. Duncan, C.T. Evelo, D. Gabriel, J. Graybeal, A. Gray, B.M. Gyori, M. Haendel, H. Harmse, N.L. Harris, I. Harrow, H.B. Hegde, A.L. Hoyt, C.T. Hoyt, D. Jiao, E. Jiménez-Ruiz, S. Jupp, H. Kim, S. Koehler, T. Liener, Q. Long, J. Malone, J.A. McLaughlin, J.A. McMurry, S. Moxon, M.C. Munoz-Torres, D. Osumi-Sutherland, J.A. Overton, B. Peters, T. Putman, N. Queralt-Rosinach, K. Shefchek, H. Solbrig, A. Thessen, T. Tudorache, N. Vasilevsky, A.H. Wagner, C.J. Mungall, A Simple Standard for Sharing Ontological Mappings (SSSOM), Database 2022 (2022) baac035. https://doi.org/10.1093/database/baac035.

[56] T. Benson, G. Grieve, Principles of FHIR, in: Principles of Health Interoperability, Springer International Publishing, Cham, 2021: pp. 79–102. https://doi.org/10.1007/978-3-030-56883-2_5.

[57] M.A. Haendel, C.G. Chute, T.D. Bennett, D.A. Eichmann, J. Guinney, W.A. Kibbe, P.R.O. Payne, E.R. Pfaff, P.N. Robinson, J.H. Saltz, H. Spratt, C. Suver, J. Wilbanks, A.B. Wilcox,

A.E. Williams, C. Wu, C. Blacketer, R.L. Bradford, J.J. Cimino, M. Clark, E.W. Colmenares, P.A. Francis, D. Gabriel, A. Graves, R. Hemadri, S.S. Hong, G. Hripcsak, D. Jiao, J.G. Klann, K. Kostka, A.M. Lee, H.P. Lehmann, L. Lingrey, R.T. Miller, M. Morris, S.N. Murphy, K. Natarajan, M.B. Palchuk, U. Sheikh, H. Solbrig, S. Visweswaran, A. Walden, K.M. Walters, G.M. Weber, X.T. Zhang, R.L. Zhu, B. Amor, A.T. Girvin, A. Manna, N. Qureshi, M.G. Kurilla, S.G. Michael, L.M. Portilla, J.L. Rutter, C.P. Austin, K.R. Gersing, the N3C Consortium, The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment, Journal of the American Medical Informatics Association 28 (2021) 427–443. https://doi.org/10.1093/jamia/ocaa196.

[58] E.R. Pfaff, A.T. Girvin, D.L. Gabriel, K. Kostka, M. Morris, M.B. Palchuk, H.P. Lehmann, B. Amor, M. Bissell, K.R. Bradwell, S. Gold, S.S. Hong, J. Loomba, A. Manna, J.A. McMurry, E. Niehaus, N. Qureshi, A. Walden, X.T. Zhang, R.L. Zhu, R.A. Moffitt, M.A. Haendel, C.G. Chute, The N3C Consortium, W.G. Adams, S. Al-Shukri, A. Anzalone, A. Baghal, T.D. Bennett, E.V. Bernstam, E.V. Bernstam, M.M. Bissell, B. Bush, T.R. Campion, V. Castro, J. Chang, D.D. Chaudhari, W. Chen, S. Chu, J.J. Cimino, K.A. Crandall, M. Crooks, S.J.D. Davies, J. DiPalazzo, D. Dorr, D. Eckrich, S.E. Eltinge, D.G. Fort, G. Golovko, S. Gupta, M.A. Haendel, J.G. Hajagos, D.A. Hanauer, B.M. Harnett, R. Horswell, N. Huang, S.G. Johnson, M. Kahn, K. Khanipov, C. Kieler, K.R. De Luzuriaga, S. Maidlow, A. Martinez, J. Mathew, J.C. McClay, G. McMahan, B. Melancon, S. Meystre, L. Miele, H. Morizono, R. Pablo, L. Patel, J. Phuong, D.J. Popham, C. Pulgarin, C. Santos, I.N. Sarkar, N. Sazo, S. Setoguchi, S. Soby, S. Surampalli, C. Suver, U.M.R. Vangala, S. Visweswaran, J. von Oehsen, K.M. Walters, L. Wiley, D.A. Williams, A. Zai, Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative, Journal of the American Medical Informatics Association (2021) ocab217. https://doi.org/10.1093/jamia/ocab217.

[59] Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality, HCUP Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses, (n.d.). https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp (accessed April 13, 2022).

[60] A. Elixhauser, C. Steiner, Most common diagnoses and procedures in US community hospitals, 1996, Healthcare Cost and Utilization Project, HCUP Research Note. Rockville, MD. Agency for Health Care Policy and Research. AHCPR Pub (1999) 99–0046. https://www.hcup-us.ahrq.gov/reports/natstats/commdx/commdx.htm (accessed April 13, 2022).

[61] P.L. Peissig, L.V. Rasmussen, R.L. Berg, J.G. Linneman, C.A. McCarty, C. Waudby, L. Chen, J.C. Denny, R.A. Wilke, J. Pathak, D. Carrell, A.N. Kho, J.B. Starren, Importance of multi-modal approaches to effectively identify cataract cases from electronic health records, J Am Med Inform Assoc 19 (2012) 225–234. https://doi.org/10.1136/amiajnl-2011-000456.

[62] S. Gold, J. Flack, TermHub, (2022). https://bit.ly/termhub.