ABSTRACT

Title of Dissertation:	VALUE SETS FOR THE ANALYSIS OF REAL-WORLD PATIENT DATA: PROBLEMS, THEORY, AND SOLUTIONS
	Sigfried Gold, Doctor of Philosophy, 2024
Dissertation directed by:	Professor Wayne Lutters, College of Information

Observational, retrospective, in silico studies based on real-world data—that is, data for research collected from sources other than randomized clinical trials—cost a minute fraction of randomized clinical trials and are essential for clinical research, pharmacoepidemiology, clinical quality measurement, health system administration, value-based care, clinical guideline compliance, and public health surveillance. They offer an alternative when randomized trials cannot provide large enough patient cohorts or patients representative of real populations in terms of comorbidities, age range, disease severity, rare conditions.

Improvements in the speed, frequency, and quality of research investigations using real-world data have accelerated with the emergence of distributed research networks based on common data models over the past ten years. Analyses of repositories of coded patient data involve data models, controlled medical vocabularies and ontologies, analytic protocols, implementations of query logic, value sets of vocabulary terms, and software platforms for developing and using these.

These studies generally rely on clinical data represented using controlled medical vocabularies and ontologies—like ICD10, SNOMED, RxNorm, CPT, and LOINC—which

catalogue and organize clinical phenomena such as conditions, treatments, and observations. Clinicians, researchers, and other medical staff collect patient data into electronic health records, registries, and claims databases with each phenomenon represented by a *code*, a concept identifier, from a medical vocabulary. Value sets are groupings of these identifiers that facilitate data collection, representation, harmonization, and analysis. Although medical vocabularies use hierarchical classification and other data structures to represent phenomena at different levels of granularity, value sets are needed for concepts that cover a number of codes.

These lists of codes representing medical terms are a common feature of the cohort, phenotype, or other variable definitions that are used to specify patients with particular clinical conditions in analytic algorithms. Developing and validating original value sets is difficult to do well; it is a relatively small but ubiquitous part of real-world data analysis, it is time-consuming, and it requires a range of clinical, terminological, and informatics expertise.

When a value set fails to match all the appropriate records or matches records that do not indicate the phenomenon of interest, study results are compromised. An inaccurate value set can lead to completely wrong study results. When value set inaccuracy causes more subtle errors in study results, conclusions may be incorrect without catching researchers' attention. One hopes in this case that the researchers will notice a problem and track it down to a value set issue. Verifying or measuring value set accuracy is difficult and costly, often impractical, sometimes impossible.

Literature recognizing the deleterious effects of value set quality on the reliability of observational research results frequently recommends public repositories where high-quality value sets for reuse can be stored, maintained, and refined by successive users. Though such repositories have been available for years and populated with hundreds or thousands of value

sets, regular reuse has not been demonstrated. Value set quality has continued to be questioned in the literature, but the value of reuse has continued to be recommended and generally accepted at face value. The hope for value set repositories has been not only for researchers to have access to expertly designed value sets but for incremental refinement, that, over time, researchers will take advantage of others' work, building on it where possible instead of repeating it, evaluating the accuracy of the value sets, and contributing their changes back to the repository.

Rather than incremental improvement or indications of value sets being vetted and validated, what we see in repositories is proliferation and clutter: new value 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 value sets for a given condition often leads value 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 value set back to the repository, they further compound the problem, especially if they neglect to document the new value set's intention and provenance.

The research offered here casts doubt on the value of reuse with currently available software and infrastructure for value set management. It is about understanding the challenges value sets present; understanding how they are made, used, and reused; and offering practice and software design recommendations to advance the ability of researchers to efficiently make or find accurate value sets for their studies, leveraging and adding to prior value set development efforts.

This required field work, and, with my advisors, I conducted a qualitative study of professionals in the field: an observational user study with the aim of understanding and characterizing normative and real-world practices in value set construction and validation, with a particular focus on how researchers use the knowledge embedded in medical terminologies and

ontologies to inform that work. I collected data through an online survey of RWD analysts and researchers interviews with a subset of survey participants, and observation of certain participants performing actual work to create value sets. We performed open coding and thematic analysis on interview and observation transcripts, interview notes, and open-ended question text from the surveys.

The requirements, recommendations, and theoretical contributions in prior literature have not been sufficient to guide the design of software that could make effective leveraging of shared value sets a reality. This dissertation presents a conceptual framework, real-world experience, and deep, detailed account of the challenges to reuse, and makes up that deficit with a high-level requirements roadmap for improved value set creation tools.

I argue, based on the evidence marshalled throughout, that there is one way to get researchers to reuse appropriate value sets or to follow best practices in determining whether a new one is absolutely needed creating their own and dedicate sufficient and appropriate effort to create them well and prepare them for reuse by others. That is, giving them software that pushes them to do these things, mostly by making it easy and obviously beneficial to do them.

I offer a start in building such software with Value Set Hub, a platform for browsing, comparing, analyzing, and authoring value sets—a tool in which the presence of multiple, sometimes redundant, value sets for the same condition strengthens rather than stymies efforts to build on the work of prior value set developers. Particular innovations include the presentation of multiple value sets on the same screen for easy comparison, the display of compared value sets in the context of vocabulary hierarchies, the integration of these analytic features and value set authoring, and value set browsing features that encourage users to review existing value sets that may be relevant to their needs.

Fitness-for-use is identified as the central challenge for value set developers and the strategies for addressing this challenge are categorized into two approaches: value-set-focused and code-focused. The concluding recommendations offer a roadmap for future work in building the next generation of value set repository platforms and authoring tools.

VALUE SETS FOR THE ANALYSIS OF REAL-WORLD PATIENT DATA: PROBLEMS, THEORY, AND SOLUTIONS

by

Sigfried Gold

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Wayne Lutters, Chair
Professor Joel Chan
Professor Jie Chen
Professor Lisa Schilling
Professor Christopher Chute

Acknowledgements

Jessica Ancker, Andrea Batch, Leilani Battle, Olivier Bodenreider, Jeff Brown, Brian Butler, Jim Cimino, Frank DeFalco, Noemie Elhadad, Niklas Elmqvist, Lee Evans, Allen Flynn, David Fram, Davera Gabriel, Beatrice Gold, Rivkah Gold, Ronna Gold, David Gotz, Erin Holve, George Hripcsak, Vojtech Huser, Guoqian Jiang, Michael Kahn, Hadi Kharrazi, Kristin Kosta, Harold Lehmann, Amanda Lazar, Linda Macri, Robert McClure, Daniella Meeker, Catherine Plaisant, Luke Rasmussen, Chrstian Reich, Rachel Richesson, Nancy Roderer, Shelley Rusincovitch, Patrick Ryan, Anthony Sena, Ben Shneiderman, Harold Solbrig, Ana Szarfman, Erica Voss, Laura Wiley, Andrew Williams, Richard Williams, Susan Winter, Meredith Zozus.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Also see Chapter 3

ADE. Adverse Drug Event.

ATLAS. The web interface to the OMOP Common Data Model and Observational Health Data Sciences and Informatics suite of software and analysis tools. https://github.com/OHDSI/Atlas/wiki.

CDM. Common Data Model.

CDW. Clinical Data Warehouse.

CMS. Centers for Medicare and Medicaid Services.

CQM. Clinical quality measure.

Clinical code set. A set of codes from specific controlled medical terminologies. Synonyms: value domain, code set, concept set.

Controlled medical terminology. A vocabulary, nomenclature, ontology, classification, or grouper authorized by a standards organization.

Code set. A set of codes from specific controlled medical terminologies. Synonyms: value set, concept set, code list, term list.

DRN. Distributed Research Network.

eCQM. Electronic Clinical Quality Measure.

EHR. Electronic health record.

ETL. Extract, Transform, Load.

HL7. Health Level 7.

i2b2. Informatics for Integrating Biology and the Bedside [99].

ICD. International Classification of Disease [163].

NCQA. National Committee for Quality Assurance [164].

NDC. National Drug Code.

NLM. National Library of Medicine.

OHDSI, Observational Health Data Sciences and Informatics. (OHDSI is pronounced "Odyssey.") A multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All its solutions are open source. OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

OMOP. Observational Medical Outcomes Partnership. Precursor to OHDSI and still used to refer to the OHDSI CDM.

PCORNet. National Patient-Centered Clinical Research Network.

Permissible value set. Permissible is a qualifier for value set or code set and their synonyms to distinguish them from phenotypic or analytic value sets. Permissible value sets are used to constrain the permissible values that can be assigned to a specific data element.

Phenotype algorithm. Synonyms: computable phenotype, electronic phenotype, or just phenotype.

Phenotypic code set. Phenotypic is a qualifier for value set or code set and their synonyms to distinguish them from permissible value set. It can be used interchangeably with Analytic. PCS are code sets used in the context of analyzing RWD. Analytic is preferred for audiences who might not be familiar with phenotype algorithms, or to highlight that their purpose is to match health records for a specific RWD analysis. Phenotypic is preferred in order to highlight that they are always components of a computable phenotype (being used in an analysis). Often a computable phenotype consists of nothing more than a single code set.

PRO. Patient reported outcome measure.

RCT. Randomized clinical trial.

RWD, Real-World Data. Data collected in digital form through the routine provision of clinical care.

RWE. Real-world evidence.

RxNorm. Drug terminology from NLM.

SNOMED CT. Systematized Nomenclature of Medicine, Clinical Terms.

UMLS. Unified Medical Language System.

Value set. The formal term denoting a set of permissible values to be used in a data element. Synonyms: value domain, code set, code list.

VSAC, Value Set Authority Center. NLM hosted code set repository.

Chapter 1. Introduction

Real-world data (RWD) [1–4], were defined by the 21st Century Cures Act as "data regarding the usage, or potential benefits or risks, of a drug derived from sources other than randomized clinical trials." [1] These data come from routinely collected sources, electronic health records (EHR), Medicare and other administrative claims, the National Death Index, tumor registries. They are the basis for observational, retrospective, in silico studies in clinical research, pharmacoepidemiology, clinical quality measurement, health system administration, value-based care, clinical guideline compliance, and public health surveillance. As the data are collected for other purposes, studies based on them cost a minute fraction of randomized clinical trials (RCT). Further, they offer an alternative when randomized trials cannot provide large enough patient cohorts or patients representative of real populations in terms of comorbidities, age range, disease severity, rare conditions.

Improvements in the speed, frequency, and quality of research investigations using RWD have accelerated with the emergence of distributed research networks (DRN) based on common data models (CDM) over the past ten years. Analyses of repositories of coded patient data involve data models, controlled medical vocabularies and ontologies, analytic protocols, implementations of query logic, value sets of vocabulary terms, and software platforms for developing and using these. The emergence of these foundational elements in computational patient research was catalyzed by new technologies in EHR use, terminology and data standards, and increasing awareness of the need for massive research data warehouses for pharmacovigilance (adverse event detection), patient safety, and clinical research.

Leaders and standards developers in in silico, observational research note that "the growing availability of [RWD] has transformed the research landscape," [5] and their importance goes beyond the drug research and regulatory applications that spurred their explosive growth a decade ago, to innumerable medical research, public health, clinical care, business intelligence, and other applications. These data and applications have necessarily grown alongside methods, tools, infrastructures, standards, and guidelines for performing and validating observational studies and other forms of research and analysis with them—particularly because they present a host of challenges not entailed in traditional RCTs. Design, execution, and validation of traditional RCTs are supported by well-established and widely accepted methods, tools, standards, and guidelines, whereas these are all rapidly evolving for RWD studies, where information models and study designs continue to be formulated and contested [6].

Though randomized clinical trials remain the standard for much evidence-based medicine, the case for RWD research is undeniable and is attested by the 574 PubMed manuscripts catalogued on the Observational Health Data Sciences and Informatics (OHDSI) [7,8] Publication Analysis webpage [9] or the 163 manuscripts listed on the National COVID Cohort Collaborative (N3C) [10] dashboard [11]. Google Scholar shows thousands of papers each for OHDSI; its predecessor organization, the Observational Medical Outcomes Partnership (OMOP); the FDA's Sentinel Initiative [3,12,13]; the National Patient-Centered Clinical Research Network (PCORNet) [14]; and N3C.

Most RWD studies rely on clinical data represented using controlled medical vocabularies and ontologies—like ICD10, SNOMED, RxNorm, CPT, and LOINC1—which catalogue and organize clinical phenomena such as conditions, treatments, and observations. Clinicians, researchers, and other medical staff collect patient data into electronic health records, registries, and claims databases with each phenomenon represented by a code, a concept identifier, from a medical vocabulary. Value sets are groupings of these identifiers that facilitate data collection, representation, harmonization, and analysis. Although medical vocabularies use hierarchical classification and other data structures to represent phenomena at different levels of granularity, value sets are needed for concepts that cover a number of codes.

These lists of codes representing medical terms are a common feature of the cohort, phenotype, or other variable definitions that are used to specify patients with particular clinical conditions in analytic algorithms. Developing and validating original value sets is difficult to do well; it is a relatively small but ubiquitous part of RWD analysis, it is time-consuming, and it requires a range of clinical, terminological, and informatics expertise.

When a value set fails to match all the appropriate records or matches records that do not indicate the phenomenon of interest, study results are compromised. An inaccurate value set can lead to completely wrong study results. When value set inaccuracy causes more subtle errors in study results, conclusions may be incorrect without catching researchers' attention. One hopes in this case that the researchers will notice a problem and track it down to a value set issue.

---

[1] These terminologies can be found on the following websites: https://www.cdc.gov/nchs/icd/icd10cm_browsertool.htm, https://www.nlm.nih.gov/healthit/snomedct/us_edition.html, https://www.nlm.nih.gov/research/umls/rxnorm, https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CPT, https://loinc.org/.

Verifying or measuring value set accuracy is difficult and costly, often impractical, sometimes impossible.

Scholarly and practical efforts to address challenges in value set management and help RWD researchers identify and select the set of codes best fitted to their hypothesis testing and analysis goals[15–23] have resulted in value set definition and documentation standards[5,24–27], methods, tools, and repositories [28–30] for authoring and sharing value sets[31–34], for assessing value set semantics and quality[35–41], and for enabling and promoting value set sharing and reuse.

Despite the development of value set repositories and other tools, concerns around value set quality persist [17,18,31,41,42]. To understand why existing tools have not solved these problems, our field needs a better understanding of the value set development process, the practices of value set authors and how those practices do or do not serve these authors.

> The utility of EMR [electronic medical records] studies has been hampered by poor quality of reporting of research methods and data. One particular area of poor reporting quality is that of clinical coding. Most EMR studies adopt bespoke definitions of clinical entities (such as disease conditions, treatments and diagnostic tests) that are seldom questioned or challenged. These clinical entities are defined through lists of 'clinical codes' and the process of preparing these code lists is rarely straightforward and often lacks rigor. … […I]t is impossible to assess the validity of the vast majority of code lists used in [RWD] research. [30]

A number of tools, methods, and strategies have been suggested for improving the quality and reporting of value sets; many of these have focused particularly on the sharing and reuse of high-quality value sets. The paper just quoted accompanied the release of Clinical Codes, a UK repository for the sharing and reuse of value sets. Around the same time, concerns about value set quality in the USA also motivated the National Library of Medicine (NLM) to develop a public repository, the Value Set Authority Center (VSAC) [28,29].

The hope for value set repositories has been not only for researchers to have access to expertly designed value sets but for incremental refinement, that, over time, researchers will take advantage of others' work, building on it where possible instead of repeating it, evaluating the accuracy of the value sets, and contributing their changes back to the repository.

Despite standardization and the availability of these platforms and other efforts to encourage value set sharing, reuse, concerns around value set quality have not abated [18,21,22,24,34,36,42–44]. Yet reuse is consistently recommended as a solution [28–30]. This study contributes to the field's understanding of value set development practices and the challenges besetting effective value set reuse [15,16,25–27,31,32,35–40]. It lays groundwork for addressing these persistent problems and challenges.

Rather than incremental improvement or indications of value sets being vetted and validated, what we see in repositories is proliferation and clutter: new value 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 value sets for a given condition often leads value 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 value set back to the repository, they further compound the problem, especially if they neglect to document the new value set's intention and provenance.

The research offered here casts doubt on the value of reuse with currently available software and infrastructure for value set management. It is about understanding the challenges value sets present; understanding how they are made, used, and reused; and offering practice and software

design recommendations to advance the ability of researchers to efficiently make or find accurate value sets for their studies, leveraging and adding to prior value set development efforts.

This required field work, and, with my advisors, I conducted a qualitative study of professionals in the field: an observational user study with the aim of understanding and characterizing normative and real-world practices in value set construction and validation, with a particular focus on how researchers use the knowledge embedded in medical terminologies and ontologies to inform that work. I collected data through an online survey of RWD analysts and researchers interviews with a subset of survey participants, and observation of certain participants performing actual work to create value sets. We performed open coding and thematic analysis on interview and observation transcripts, interview notes, and open-ended question text from the surveys.

The requirements, recommendations, and theoretical contributions in prior literature have not been sufficient to guide the design of software that could make effective leveraging of shared value sets a reality. This dissertation presents a conceptual framework, real-world experience, and a deep, detailed account of the challenges to reuse, addressing that deficit with a high-level requirements roadmap for improved value set creation tools.

I argue, based on the evidence marshalled throughout, that there is one way to get researchers to reuse appropriate value sets or to follow best practices in determining whether a new one is absolutely needed before creating their own and, if so, in dedicating sufficient and appropriate effort to create them well and prepare them for reuse by others. That way is by giving them software that pushes them to do these things, mostly by making it easy and obviously beneficial to do so.

I offer a start in building such software with Value Set Hub (VS-Hub), a platform for browsing, comparing, analyzing, and authoring value sets—a tool in which the presence of multiple, sometimes redundant, value sets for the same condition strengthens rather than stymies efforts to build on the work of prior value set developers. Particular innovations include the presentation of multiple value sets on the same screen for easy comparison, the display of compared value sets in the context of vocabulary hierarchies, the integration of these analytic features and value set authoring, and value set browsing features that encourage users to review existing value sets that may be relevant to their needs.

This introductory chapter gives a preliminary explanation of the basic ideas and historical context of this work and the arc of the remaining chapters. To avoid redundancy, it will provide only a high-level account of what value sets are, the problems they present, and the contexts surrounding them. More background is given in the introductions to individual chapters.

1.1 Historical context and audience

Value sets play various roles in EHRs, data capture, billing, clinical quality, and health analytics. Health analytics is the focus of this research. Value sets have received more attention in literature about other contexts they are used in than in RWD research, in which they are subsidiary to phenotype or cohort algorithms. They are essential for determining the patient records processed by phenotypes and study algorithms and very strongly affect study results. Part of the reason their use in RWD research receives little attention is that they are structurally much simpler than their containing algorithmic objects. This dissertation makes a case for the importance of high-quality analytic value sets and explores in unprecedented detail the challenges involved in developing them.

This research directly participates in and addresses the small literature around value set use in RWD applications. It should be of wider interest to the larger communities of informaticists working in medical terminologies, patient safety, and RWD research generally. And it should be of particular interest to researchers, informaticists, and IT professionals who support the infrastructure for distributed research networks (DRN) based on common data models (CDM): OMOP/OHDSI, PCORnet, All of Us [45], and the N3C.

The work presented in this dissertation emerged from a specific historical context that made it possible and necessary. It will be of immediate interest to the small but crucial cadre of informatics professionals who provide software and infrastructure (e.g., VSAC, OHDSI, N3C) or theory and instruction around research uses of medical terminologies and value sets [46,47,43,30,17] or have otherwise been brought to give them any considerable thought. This dissertation fits firmly in a relatively new niche of clinical research informatics focusing on RWD, CDMs, and clinical research networks; it takes up the concerns of scholars in that area and primarily addresses them. But I came to this study from work in the temporal summarization and visualization of clinical data, specifically of RWD, where exciting things were happening between 2010 and 2015. Innovation and continued work in this area appears to have slowed since then. For me, like some others[2], I took what I originally thought would be a detour into terminology and value sets from working on more exciting aspects of RWD analytics, having realized that these analytic projects would be impossible without better ways of capturing, standardizing, and categorizing clinical meaning.

[2] For instance, two of the founding practical and theoretical contributors to contemporary terminology use in RWD research, Christopher Chute and James Cimino, cited just above.

1.1.1 An historic transition in RWD research

Although clinical research based on EHR, registry, and reimbursement claims data has been ongoing for many decades, the advent of DRNs utilizing RWD and CDMs has significantly transformed and accelerated the field (Table 1). This transformation initially took root in post-marketing surveillance, pharmacovigilance, and pharmacoepidemiology. The catalyst for massively expanding the availability and utility of RWD in the USA was the Vioxx debacle in the late nineties, which underscored the critical need for extensive and diverse data sources. In

Table 1. Milestones in real-world data analysis

MILESTONE	EXAMPLES
1850s – Systematic data collection for analysis	Florence Nightingale inaugurated modern epidemiology, systematically collecting and analyzing data on Crimean War soldiers' morbidity and mortality, demonstrating the impact of sanitary conditions on health outcomes [49].
1920s – RWD for research	Early case control and longitudinal studies using routinely collected medical data: breast cancer study starting in 1926 [50], Framingham Heart Study starting in 1948 [51].
1960s – Digital databases for primary and secondary use	The MUMPS programming language at Veterans Administration hospitals and the ARAMIS Database for Rheumatology the creation and management of electronic health records and the analysis of patient data. [52,53]
1980s – Massive collection of medical records for research purposes	The General Practice Research Database (GPRD) has collected anonymized patient data from general practitioners in the UK since the 1980s and continues to be used extensively for epidemiological research and pharmacoepidemiology. [54]
1990s – Foundational work in the theory and practice of terminology use for data collection and analysis	Desiderata for controlled medical vocabularies in the twenty-first century [44] and The Copernican era of healthcare terminology: a re-centering of health information systems [45]
2000s – DRNs for conducting large, cross-site RWD studies based on common study algorithms, phenotype definitions, and value sets	The Electronic Medical Records and Genomics (eMERGE) Network (2007): The eMERGE Network was established to link electronic medical records (EMRs) with genomic data to enable genome-wide association studies (GWAS). This initiative has significantly advanced personalized medicine and the understanding of genetic factors in disease.
2010s – CDMs for harmonizing analytic code in research networks	With eMERGE, each site implemented the study and phenotype algorithms for their own data warehouses. Starting around 2008, federated DRNs based on common data models and harmonized value sets made it possible to replicate studies using the same software code. These included OMOP/OHDSI, Sentinel, and PCORnet.
2020s – Centralized research networks	All of Us and N3C made it possible not just for the same code to be executed across different sites, but to analyze data from different sites on a central, harmonized data warehouse platform.

2007, legislative and regulatory frameworks, the Prescription Drug User Fee Act (PDUFA IV) and the FDA Amendments Act (FDAAA), facilitated this expansion by mandating more rigorous post-market surveillance with programs like the Sentinel Initiative, OMOP [48,49], and PCORnet [50]. The early OMOP meetings aimed to produce a common data model following the 80/20 work effort principle, although in practice, the reduction was even more extreme. EHRs often have thousands of tables and tens of thousands of columns, which OMOP harmonizes down to a handful of primary tables with dozens of columns. This simplification enabled the development of analytic software applicable to most clinical conditions, usable by any institution adhering to the OMOP model. As methods, tools, and research networks employing CDMs developed, particularly within the OHDSI community, these new observational research methodologies expanded their influence beyond drug safety into broader clinical research applications. This was exemplified in early 2020 when the N3C rapidly mobilized to study COVID-19, leveraging a vast network of data partners and researchers.

In 2008, I worked at PhaseForward with the Army Pharmacovigilance Center at the forefront of developing innovative software solutions for leveraging EHR and claims data to enhance pharmacovigilance and post-marketing surveillance. This period marked a significant shift from reliance on traditional spontaneous reporting systems like the FDA's Adverse Event Reporting System (AERS) towards the utilization of RWD, which offered a broader and more dynamic perspective on patient outcomes and drug safety.

The OHDSI initiative and the OMOP CDM grew out of this pharmacoepidemiology work and played pivotal roles in standardizing and harmonizing diverse health data sources. These frameworks provided a robust infrastructure that enabled the harmonization and analysis of

RWD on a global scale, thus facilitating more rapid and cost-effective clinical research. The implementation of these standardized models significantly enhanced the interoperability of health data, enabling researchers to conduct collaborative and reproducible studies across multiple institutions.

This reduction in complexity was crucial for forming large international DRNs, allowing replication of studies by simply sharing and executing study algorithms across data warehouses. OMOP not only simplified data structures but also harmonized data semantics by choosing standard vocabularies (such as SNOMED, RxNorm, and LOINC) and mapping codes to them from source vocabularies (such as ICD, Read, MeSH, MedDRA, NDC). While this harmonization provided massive data resources and facilitated software design, it did entail information loss. Moreover, existing longitudinal analysis methods designed for clinical trials could not be directly applied to RWD, as data is captured only when patients seek treatment.

1.1.2 Temporal summarization and visualization of RWD

As new and existing observational analysis methods were applied, it became apparent that researchers were obtaining results that were often perplexing. For example, known effects, such as the correlation between ACE inhibitors and increased incidence of angioedema, were not consistently detected by statistical tests. This prompted a focus on temporal patient visualizations, initially at the level of individual patient trajectories and subsequently for visualizing event sequence patterns across patient groups. While these visualizations helped clarify some of the unexpected results from statistical analyses, their practical application was limited to relatively small patient cohorts, typically a few hundred at a time.

At the University of Maryland's Human-Computer Interaction Lab (HCIL), researchers like Ben Shneiderman, Catherine Plaisant and their students (particularly Krist Wongsuphasawat and Megan Monroe) were pioneering temporal pattern visualization of medical data with tools like Lifelines and LifeFlow. LifeFlow was limited to visualizing sequences of point events. For drug safety research, it was essential to account for periods of drug exposure, not just the initiation of drug use. During my tenure at Oracle's Health Sciences Unit, I facilitated an Oracle research grant to HCIL, supporting development of EventFlow, a version of LifeFlow that could handle event interval data.

EventFlow was designed to facilitate a clearer understanding of patient trajectories by providing an intuitive visualization interface. However, the broader adoption of such tools encountered significant challenges. EventFlow and other temporal pattern summarization visualizations struggle to effectively communicate patterns involving more than around seven event types. Given that RWD concepts number in the tens of thousands, it became evident that filtering and aggregating concepts were necessary steps before aggregating patient records by concept group.

I initially made the mistake many others do of assuming, since medical vocabularies are hierarchical, it should be possible to analyze temporal event correlation for pharmacovigilance signal detection and other RWD analysis problems by aggregating events at appropriate levels of granularity. This would have to be determined experimentally: finding a level of granularity coarse enough not to lose signal to underpowered cohort sizes and fine enough not to lose signal to noise from events that didn't cause the adverse event. Whether by computational methods or

manual exploration by an expert in the relevant clinical area, one would aggregate concepts using vocabulary hierarchies, starting at a low granularity and progressively drilling down.

Though it is certainly helpful to drill through hierarchies to select particular codes and their descendants, it is very frequently insufficient for finding sets of concepts that would group patient records into useful, clinically meaningful categories. A given subtree of concepts, despite all having an *is-a* relationship to the subtree's root, may include concepts that do not actually indicate that the patient has experienced the phenomenon of interest or may not include concepts that do. That is why value sets are needed.

Unfortunately, however, high-quality value sets are not easy to make. Worse, there is no particular best value set for finding a single clinical phenomenon like type 2 diabetes mellitus in a database; it depends on many contextual factors. However, value sets allow more flexibility than single codes and their descendants.

Communities like OHDSI and N3C have already transformed the way medical research is done over the past decade. I believe temporal summary visualization like EventFlow would greatly accelerate the quality and efficiency of evidence creation in these communities, but such tools will be useless without significant progress in methodologies and tools for clinical event classification and aggregation. The problems around value set quality and reusability are far from being solved, but this dissertation brings us closer.

Even without progress sufficient to allow effective temporal pattern recognition and massively aggregated event sequence visualization, thousands of studies have been published on the basis of recent developments in RWD analysis, and the quality and efficiency this work will

progress as researchers' ability to make or reuse accurate value sets progresses, ultimately contributing to the advancement of medical research and patient care.

1.1.3 Shifting focus from software development to user research and back

The goal motivating all of the research presented here has been the development of software to improve the quality and efficiency of RWD analysis. Over the previous decade or so, I had spent countless hours in meetings with epidemiologists, clinicians, informaticists, biostatisticians, and programmers developing and implementing protocols and algorithms for studying drug-adverse event associations, which frequently involved choosing drug or clinical event codes. I was a programmer myself but saw first-hand the frustrations these experts experienced selecting codes and making sense of patient data, sparking numerous feature ideas and design plans, mostly involving visualization of patient trajectories and terminology hierarchies. Some of these I implemented (and remain in use years later), others were more ambitious, beyond what my employers were willing to fund. In my efforts to realize these plans, I focused my work life on organizations in the OHDSI community and I began doctoral study.

Prior to beginning the research of this dissertation, I had been working on visualization software for patient data analysis, temporal pattern discovery, and grouping patient records by clinical conditions and treatments. The focus of that software increasingly narrowed to terminology exploration, selection, and grouping, both for the reasons given above and because there were no sources of patient-level data available to me at UMD. Even after putting aside efforts to integrate patient data and terminology features, I have had an evolving set of design plans that would be positive contributions to the quality and efficiency of RWD analysis. Yet, the scope of these plans and the set of problems they are meant to address has been more than I

could accomplish with available resources. I had to go beyond my own knowledge and experience and learn what practitioners in the field needed in order to narrow and prioritize requirements and design proposals.

Since then, it has become clear that the problems experienced in value set management were greater than I (and others in the field) had realized and that exploratory and theoretical work to understand them would itself produce new and important contributions to the literature and practice of observational research informatics.

Despite success and breaking new ground in these efforts, they would not lead me to a set of prioritized problems and features small enough to be implemented and evaluated over the course of a single PhD. In 2021, I started work for N3C, bringing me to one of the largest communities of RWD researchers working with the same tools on the same (harmonized and centralized) warehouse of real-world data ever assembled. There I was able to combine what I learned through this research with the immediate needs of this large community into the requirements and implementation of VS-Hub, the software discussed in Chapter 6.

1.2 Arc of the research

This dissertation consists of one published paper, two papers currently under review, one unpublished paper, two chapters integrating material from other papers (finished, unfinished, submitted, and unsubmitted), as well as this introduction and a concluding chapter. (I was the primary author of the papers included here; coauthors, where listed, contributed mostly through conversations, document comments, and light copy editing).

Looking over the entire arc of our studies, we have been motivated by these central research questions

- Why, after much work and scholarship around value sets and value set repositories for encouraging reuse, does it seem as though reuse is relatively rare, at least in RWD contexts?
- How can we reduce confusion over what value sets are and the nomenclature surrounding them?
- What tools and practices are common in value set development, use, and reuse?
- How should we formulate conceptual and process models to understand and describe these practices?
- What barriers are faced by developers and re-users of value sets in RWD research?
- How can our descriptive models also serve prescriptively to inform better practices?
- What software functionality can we design and implement based on our findings and models to improve value set authoring by leveraging previously created value sets?

1.2.1 Chapter 2: The problem space

Chapter 2 is based on a paper published in the proceedings of the 2018 AMIA Symposium: "Clinical concept value sets and interoperability in health data analytics." [51] It is a theoretical examination of value sets in RWD research, based on the professional experience of myself and several expert coauthors, discussing the importance of value sets and their place in health data analytics, as a linchpin of interoperability. It associates their role in interoperability with the burgeoning growth of distributed research networks based on common data models. CDMs are not the only contexts in which value sets and phenotype algorithms are particularly important. They are just as central in any analytic use of RWD. But CDMs and shared software stacks make the possibility of sharing phenotypes much more immediate. With these platforms, observational study code can be shared and run immediately across sites.

Although my ideas have evolved in many ways since this publication, I still stand by the arguments and recommendations made in it. Beyond my own professional experience, it is based on conversations and informal interviews with about 15 leading informatics and terminology experts, eight of whom served as co-authors. They corrected my misapprehensions, educated me

where gaps in my understanding damaged the argument, and verified the soundness of the final product.

The paper focuses on reuse as a primary goal, but as a more problematic goal than previous authors in this space had acknowledged.

1.2.2 Chapter 3: Defining terms

Along with colleagues deeply immersed in the field of RWD analytics, I wrote a small paper meant to disambiguate and clarify a few essential terms in the form of a citable glossary that could be used across research, regulatory, and health administration communities.[3] Chapter 3 and the glossary in **Error! Reference source not found.** can help in keeping track of the central objects, terms, and acronyms used throughout.

1.2.3 Chapters 4 and 5: Empirical investigation, theory, and conceptual/process models

The contributions and recommendations in Chapter 2, sound as they may be, were theoretical and based on coauthor experience rather than specific evidence. Though aware of many goals, use cases, and challenges people faced when composing value sets, I didn't know which of these were most common or which features could benefit which communities, or which, if any, communities or subcommunities or individuals were aware of the shortcomings in their available tools for value set management or experienced these shortcomings as problems that might be worth solving. The informatics literature on value sets seemed to have nothing to say on these

[3] The four of us also collaborated on "Electronic health records–based phenotyping [19]," a 2020 chapter in an NIH publication, "Rethinking clinical trials: A living textbook of pragmatic clinical trials." That chapter is not included in the dissertation as writing was shared equally amongst the coauthors.

matters. The literature seemed to consist exclusively of solutions to particular problems and evidence that these solutions could advance the field in solving these problems, but the importance and scope of the problems was not substantiated with much evidence.

As we have said, sharing and reuse are often mentioned as vital to addressing value set quality issues, and the repositories have been built, but no one is claiming that they are actually fostering reuse. Chapter 2 contributed further thought on the difficulties with value sets and their reuse, but did it correctly identify the barriers? Was reuse happening or not? Was it needed? By whom? That paper tried to establish that it could not happen without much more incentive and powerful tools. Was that the case? And could the same tools be used for everyone making value sets?

My advisors and I decided to go deeper and conduct field work: a qualitative study of professionals in the field: a synthetic literature review and observational user study with the aim of understanding and characterizing normative and real-world practices in value set construction and validation, with a particular focus on how researchers use the knowledge embedded in medical terminologies and ontologies to inform that work.

We designed and conducted a qualitative study, including an online survey to help identify RWD analysts and researchers; interviews with a subset of survey participants; and observation where possible of participants performing actual work to create value sets. We performed open coding and thematic analysis on interview transcripts, interview notes, and open-ended question text from the surveys.

The specific research questions that motivated this study were:

- What resources, practices, tools, and infrastructures are used in managing phenotypic value sets?

- Does a different picture emerge when we look across RWD analysis contexts and use cases—not just formal, observational research studies for publication but for exploratory, ad hoc, or other internal analyses? Do we see patterns in uses, motivations, challenges, even across this wide spectrum?
- How are vocabulary and other semantic resources used in electronic phenotyping and other analytic use of RWD?

The results revealed a set of disparate perspectives and practices—in context, validation methods, and basic nomenclature[4]—that further complicate possibilities for reuse.

1.2.3.1 Real-world analysis of the analysis of real-world data

Chapter 4 contains the bulk of the results and findings of our study. It combines material produced over the course of about three years, including parts of papers submitted to qualify for PhD candidacy and a preprint, "Practices, norms, and aspirations regarding the construction, validation, and reuse of code sets in the analysis of real-world data [20]."

In response to the rejection of that preprinted paper by *The Journal of Biomedical Informatics* (JBI), I went through many iterations of clarification, refinement, and re-vision [52]. In addressing the comments of reviewers, coauthors, and others who generously read it, my own thinking evolved, and my conceptual and process models became much clearer, resulting in an essentially new paper. On the verge of submitting the new paper to JBI, I had the almost comical experience of learning that they no longer accepted submissions of any length and that their new length requirements (still more generous than other informatics journals) meant cutting it by a third. Some of the cut material, including the CEER model (Section 4.6), is also integrated into Chapter 4.

[4] This provided the impetus for the Chapter 3 paper, which was written concurrently with writing up the study.

The need to shorten and consolidate (along with my experience working on N3C) led to my recognizing a single, distinct problem as the kernel of my study results and findings.

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

Throughout the work integrated into Chapter 4, redundancy appeared as a significant though not central problem. Prior to the field study, though, I was barely aware of its importance. In Chapter 2, I even argued for cooperation, consolidation, or centralization of value set repositories in order to generate positive network effects. The idea was that the more people work in a single repository, the more they will improve the value sets in that repository. What actually happens is that as value set repositories gain wider use, they accumulate redundant and low-quality value sets, making it increasingly difficult for a potential re-user to identify high-quality value sets appropriate to their needs. This is the problem that needs to be solved before value set repositories can be widely used and useful. Positive network effects will only accrue if all contributions to a repository are dedicated either to improving existing value sets or making new ones when absolutely necessary. Redundancy is the focus of Chapter 5.

1.2.4 VS-Hub: Software implemented to solve redundancy and authoring problems

The understanding I gained in writing the papers already described, prepared me well to work on the N3C team that supports researchers in using and contributing to N3C's repository of almost 8,000 value sets. In this role I was able to design and implement software and return to the passion with which I began my doctoral studies: interactive visualization for complex terminology hierarchies and multiple collections of value sets.

Chapter 6 presents VS-Hub, a platform for browsing, comparing, analyzing, and authoring value sets—a tool in which the presence of multiple, sometimes redundant, value sets for the same condition strengthens rather than stymies efforts to build on the work of prior value set developers. Over a five-month period, VS-Hub has been used by over 200 users and has been used in the development and curation of 95 recommended value sets for commonly studied conditions, treatments, and lab tests. Particular innovations include the presentation of multiple value sets on the same screen for easy comparison, the display of compared value sets in the context of vocabulary hierarchies, the integration of these analytic features and value set authoring, and value set browsing features that encourage users to review existing value sets that may be relevant to their needs.

1.2.5 Conclusion

Chapter 7 presents a new synthesis of the findings of the preceding chapters. Fitness-for-use is identified as the central challenge for value set developers and the strategies for addressing this challenge are categorized into two approaches: value-set-focused and code-focused. The recommendations given in Chapter 7 clarify and expand on those in the earlier chapters and offer a roadmap for future work in building the next generation of value set repository platforms and authoring tools.

Chapter 2. The problem space: Clinical concept value sets and interoperability in health data analytics[5]

This paper focuses on *value sets* [36,43,53–55] as an essential component in the health analytics ecosystem.[56–61] While value sets appear in many contexts and serve many purposes and HL7® offers a general specification,[25] our discussion and recommendations focus only on the analytic context; i.e., a value set defines a collection of codes or terms from controlled medical vocabularies that are treated as equivalent for use in a clinical query or analytic task. In order for a clinical idea used in an analytic task to be applied to coded patient data, it will be associated with a collection of concepts—represented as codes from code systems, i.e., a value set—that, when taken as a uniform collection, can be used in identifying a cohort of patients or set of patient records matching that idea. A patient cohort is identified by finding value set member codes in patient data records. For instance, a value set representing ACE inhibitor exposure might include thousands of NDC and RxNorm codes. A query using a well-constructed value set of ACE inhibitors that is appropriate for the query context, should return results (e.g., 30 Lisinopril 40 MG Oral Tablet dispensed to patient 123 on 1/1/2011) of the highest possible relevance and recall.

The phrase *value set* is problematic. Our usage may be confusing to those familiar with value sets as criteria for populating drop down lists or for constraining the values allowed in a data element. The term may also be unfamiliar to health data researchers and analysts who routinely

[5] Published as "Clinical concept value sets and interoperability in health data analytics," AMIA Annual Symposium Proceedings 2018.

construct value sets to query encoded data but call them by a different name (e.g., code lists or concept sets) or may not recognize them as distinct components of analytic algorithms at all.

We will discuss shared repositories of reusable value sets, some of which are already in use, and offer recommendations for their further development and adoption. In order to motivate these contributions, we explain (1) how value sets fit into specific analytic tasks and the health analytics landscape more broadly; (2) their growing importance and ubiquity with the advent of Common Data Models, Distributed Research Networks, and the availability of higher order, reusable analytic resources like electronic phenotypes and electronic clinical quality measures;[62] (3) the formidable barriers to value set reuse; and (4) our introduction of a concept-agnostic orientation to vocabulary collections. The costs of ad hoc value set management and the benefits of value set reuse are described or implied throughout. Our (5) standards, infrastructure, and design recommendations address are not systematic or comprehensive but invite further work to support value set reuse for health analytics.

2.1 Value Sets in Health Analytics

We confine our discussion of clinical research and health analytics to contexts in which the data has been collected already in the process of providing care, i.e., *secondary use.* This excludes much clinical research—randomized control trials, prospective cohort studies—but makes the discussion relevant to important, non-research secondary uses of health data such as health administration, health economics, public health surveillance, which involve similar processes, resources, and challenges. Secondary use analyses, by definition, depend on data collected without regard for their analytic goals and often lack variables and observations central to their questions. At the same time, they can leverage datasets orders of magnitude larger than

the expenses of randomized clinical trials would allow, accelerating formulation, execution, and reformulation of questions with a flexibility and speed impossible in human subjects research. Given our focus on value sets, we further confine our scope to data encoded with controlled medical vocabularies, ignoring narrative text and complex objects like lab results and images.

Figure 1 schematizes the life of a health data analytics task as a process: (1) formulation of a question; (2) selection of a method; (3) selection of a software implementation of that method; (4) execution on data with appropriate parameter configuration; and further steps in which the results may prompt more analysis, be shared with DRN collaborators, or be used in publications or reports, to address patient needs, or otherwise disseminated.

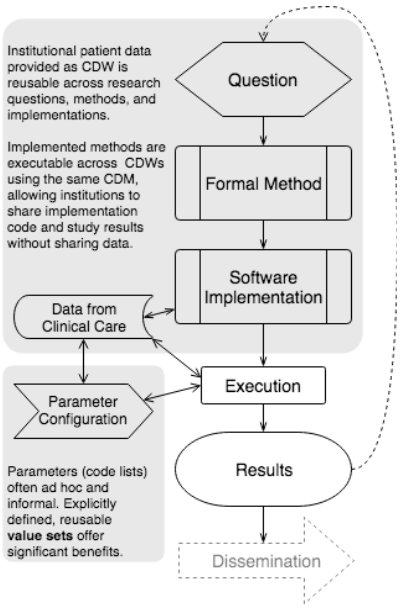

Figure 1. Abstract model of health data analytics

Obviously, the generation and capture of data by patients and care providers is the substrate on which the execution engine will run. A substrate directly influenced by a vast ecosystem of terminology standards, data transmission standards, networks, software, government policies, regulatory agencies, funding agencies, and health systems, not to mention the IT services and infrastructures of the institution where the data analysis occurs.

Most questions can be addressed using well-understood methods (from e.g., statistics, epidemiology, health economics), and any widely used method will, of course, be applicable to an unbounded set of questions. Formal methods, further, can be implemented in countless ways: through guided interaction in specialized applications, with predefined functions in statistical packages and other analysis tools, or coded ad hoc in generalized programming platforms.

However the method is implemented, its execution will require connection to some sort of CDW. In selecting or developing both methods and implementations, analysts should prefer those that are already established and validated if they are available and appropriate. Developing new ones is time-consuming, error-prone, and complicates interpretation of results and comparing them with results from similar analyses.

As shown in Figure 2, regardless of the method and implementation chosen, method inputs (like treatment or outcome) must be specified with particular clinical concepts (lisinopril, angioedema), which should be expressed as value sets, i.e., collections of codes matching relevant records in the data. There are considerable advantages to expressing clinical concepts with established, preferably validated, value sets. The clinical abstractions needed for analysis are better understood and clarified when expressed as value sets because the fitness for the subject matter of the question (specific diseases, treatments) must be starkly examined when discretely specified. Having done so, value sets can be reused across questions, inferences combining results of diverse studies of particular topics are more likely to be meaningful.([36] Surprisingly then, examination of the work in the areas of research informatics and semantic interoperability suggests that value set reuse is more the exception than the rule. It occurs when mandated (as with eCQMs for clinical quality accreditation) or for value

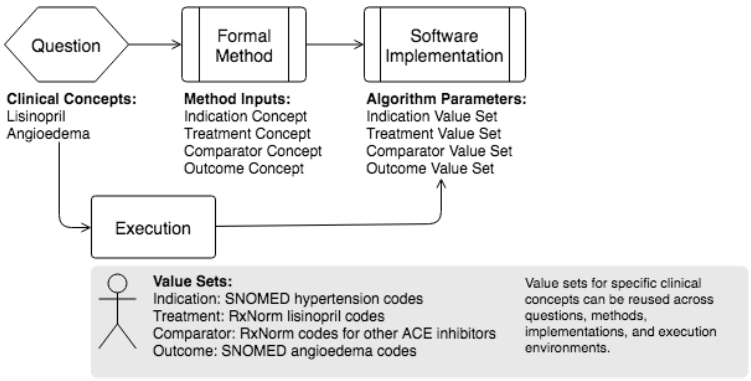

Figure 2. Value sets represent question concepts as parameters for method execution.

- **Analyst, as researcher, asks, "Does lisinopril cause angioedema?"**
- **As biostatistician, chooses a formal effect estimation method.**
- **As data scientist or programmer, chooses or constructs implementation that require method input be specified as value sets.**
- **As informaticist or end user, chooses or constructs value sets to representclinical concepts for execution.**

sets that happen to be configured in terminology services that happen to be attached to query interfaces used in analysis, which is not the norm.

A value set is an enumerated list or set of selection criteria that resolves to an enumerated list of codes or terms appropriate to a coded data element. Comprised of a versioned *value set definition* that, when applied against code systems, generates a set of usable codes known as the *value set expansion*. Unlike software implementations of analytic methods, value sets do not need to be expressed in specific programming or query languages. Most simply, they can be expressed as enumerated code lists, in which case a value set's expansion may be identical to its definition. An example of a value set definition expressed as a set of selection criteria could be: all the drug product descendants of a particular concept (ACE inhibitors) with the exception of drug products containing a particular ingredient (lisinopril). There is not yet a single widely accepted grammar for these rules, but a suitable one could be programming language independent. Even when a value set is defined through rules rather than as an enumeration of codes, it must be resolved into an enumeration of codes before being passed as a parameter to an analytic method implementation. The development and curation of value sets can be managed independently from the objects that depend on them.

Although recognition of the difficulty and benefits of high quality value set development and reuse is not new [63–67], as evidenced by projects such as NLM's VSAC, we believe the time is ripe for renewed attention and efforts at cross-domain collaboration. VSAC offers a central hub for value set curation in the clinical quality measurement community (CMS, NCQA), but a burgeoning array of secondary use analytics projects, resources, and networks (e.g., All of Us, OHDSI, Sentinel, PCORNet, i2b2, commercial analytics products from EHR vendors and

clinical data aggregators) have gone their own way. Secondary analytics tools, for the most part, have tightly coupled value set management with cohort definition and other query capabilities and users seldom share value sets even in when using a single tool at a single institution, much less across tools, institutions, and communities. The tool developers cannot be faulted for taking this approach (or non-approach, as it were). The astounding advances made in DRN development and infrastructure and resources for secondary health analytics would not have been possible if developers had tied their value set definition functionality to external technologies, services, and standards.

2.2 Common Data Models

The emergence of CDMs over the past decade has made interoperable analytics possible in the drug safety and clinical research communities that have adopted them. [48,49,58,59,68–75] These data models have evolved rapidly as a result of the opportunities they offer for study reproducibility, observational methods development, tool reuse, and coordination of research across diverse institutions without the need for patient-level data sharing. Software and system infrastructures have sprung up around CDMs in support of their use, encompassing platforms that extend existing, well-established informatics infrastructures, and creating a network effect of exponentially increasing benefits as their adoption spreads.[76,77]

CDMs allow clinical data networks to share queries, observational study methods, and analytic code. Syntactic interoperability results from sharing a common database schema and standardizing database engine support to allow queries and code to run without generating errors. CDMs must also provide for semantic interoperability by standardizing their use of semantic

resources so that query results have compatible meanings across application to different data repositories.

The predominant public CDMs are (in order of their inception) i2b2, OHDSI (originally called OMOP), Sentinel, and PCORNet. The existence of multiple CDMs can be confusing for potential adopters. Efforts at harmonizing them are being made, [78] but leadership of the CDM organizations are divided by philosophical differences and the different needs of their primary stakeholders, not to mention organizational rivalries.

Those of us active in the Observational Health Data Sciences and Informatics (OHDSI) community have a distinctive perspective on value sets (called "concept sets" in that community) as OHDSI's vocabulary system includes a multiplicity of vocabularies in each of several domains: e.g., ICD9, ICD10, SNOMED CT®, Read, for conditions; NDC, RxNorm, ATC, for drugs. Integrated vocabulary collections allow the CDM, analytic framework, and study code to be shared amongst DRN members using diverse source systems and vocabulary encodings. OHDSI, like UMLS, maps each code or concept from all of their constituent vocabularies to a single, authoritative concept in the collection; in UMLS every distinct unit of meaning is unambiguously associated with a CUI or Concept Unique Identifier; in OHDSI's vocabulary system, codes or concepts from particular, robust vocabularies or ontologies (e.g., SNOMED CT, RxNorm) are tagged as "standard" or "target" concepts, and items from other vocabularies are mapped to these. Users converting data to the CDM must associate each record with an appropriate standard concept, as well as retaining a reference to the original concept the source data was coded with.

In addition to supporting query reuse across data encoded with diverse vocabularies, integrated vocabulary collections also allow semantic information embedded in them to be leveraged in code selection. As an example, DeFalco, et al. [79] takes terms from three different drug classification vocabularies (ATC, NDF-RT, and ETC) and follows mappings to three overlapping sets of NDC codes, which they combine into a single value set they use to represent opioid exposure.

OHDSI's strategy for achieving semantic interoperability is not without critics. Sentinel's CDM requires that clinical codes be represented and queried in their original encodings [80] to prevent information loss and ambiguity. This can work for Sentinel, which is a centrally controlled DRN, has specific mandates and funding, and has contractual relationships with its DRN members and can require them to use approved code systems and meet rigorous data quality measures. OHDSI, on the other hand, is a voluntary, open collaborative and DRN, bound together by its CDM, a large set of interconnected open-source software tools, and an active community of contributors and users.

OHDSI's rapid growth—in user base, user diversity, and technical platform—has led its Architecture Workgroup [81] to begin developing formal OpenAPI specifications for value sets and cohort definitions. This puts OHDSI at a critical juncture: it can take this opportunity to engage the wider informatics community and align with those approaching the same problems in different contexts, or risk reinventing standards and technologies and complicating future cross-domain collaboration. OHDSI's confrontation with value set specifications will be of interest to a wider audience because OHDSI faces challenges that other efforts have and will continue to face

in this arena, as well as facing challenges involved in its international user base and its need to support a wide array of redundant or overlapping vocabularies.

2.3 Barriers against reuse of value sets

Standards for content and structure, platforms for development and maintenance, and repositories for value set sharing already exist, though many of the benefits of reuse are not yet realized. Even with platforms and repositories that make value set sharing technically possible, practices that would lead to reuse are not in place. For researchers or analysts who need a value set to represent some clinical concept in the context of developing a cohort definition or quality measure, the tendency is to create their own rather than taking the trouble to find an existing value set for that concept and verify that it meets their needs.

As an illustration, Organization A and Organization B belong to a DRN, use the same CDM, and use a common repository of value sets. Org. A defines ACE inhibitors as a particular list of RxNorm or NDC codes for use in a cohort study. Org A'a value set may be syntactically and semantically interoperable, i.e., *technically reusable* such that Org. B could use it for a new study involving ACE inhibitors, and it will work in their environment on their data as expected. But this reusability is a far cry from *real-world reuse*. For Org. B to actually reuse Org. A's value set would require: 1) that they can find it; 2) that they believe it's worth the effort to find it rather than defining a new one; 3) they can verify that it serves their current purpose; 4) if it doesn't quite, then Org. B, as a contributing member of a value set reuse community would modify it accordingly and document their change in an easily auditable way so potential future users would understand the difference and, in turn, use or modify the version closest to their own needs.

In a well-used value set repository, common clinical concepts are likely to have many variant value sets, differing in possibly subtle ways to capture certain use cases or clinical nuances. For this reason, finding the most appropriate match for the analyst's immediate task may prove time consuming. The logical complexities involved in crafting cohort definitions and other analytics are rife with technical and cognitive challenges. In allocating cognitive resource, the chore of code selection is unlikely to receive more than the minimum attention necessary. Even if a conscientious analyst determines that creating or revising a value set is necessary, allowing for reuse will burden her with the extra work of adding her new value set to the repository, documenting, and naming it, with no guarantee that this work will benefit anyone else. In certain cases, a quick text search or vocabulary perusal may yield a perfect value set for a given purpose. Creating one-off value sets without worrying about reuse allows the analyst to format codes to match her data and to render her value set directly as a filtering criterion in the query where it's needed; no need for translation, data type conversion, joins to vocabulary tables, or consideration of vocabulary versions.

The disincentives for reuse practices in analytic workflows are immediately felt, while benefits may be unclear, uncertain, or only available to future users.

One place shared value sets are currently being used is for electronic clinical quality measures (eCQM). The eCQM "Statin Therapy for the Prevention and Treatment of Cardiovascular Disease" from the eCQI Resource Center[62] is a multi-step algorithm making reference to numerous clinical concepts whose definitions are in the form of value sets specified remotely by the NLM's VSAC. The VSAC, in combination with the functionality provided by JIRA commenting and the companion NLM VSAC Collaboration site, is designed to create and then

improve high-quality value sets through reuse and refinement, in addition to supporting distribution of specific value sets for compliance with CMS requirements. The capabilities NLM's tools provide is only a starting point to address the difficulty practical semantic interoperability faces.

2.4 A Concept-Agnostic Perspective on Terminology Systems

No in-depth encounter with value sets and terminology systems can entirely avoid dealing with certain semiotic and ontological difficulties. Jim Cimino's foundational 1998 and 2006 desiderata papers [46,82] establish norms and language that would suffice if it were not for the need to consider value sets that draw from overlapping vocabularies. Cimino's "concept orientation" desideratum calls for nonvague, nonambiguous, and nonredundant vocabularies that classify their domains into clear divisions and subdivisions. Concepts are the fundamental units of meaning in such vocabularies, unlike terms, labels, or synonyms, which are names used to denote these concepts, to convey their meaning. We introduce the idea of a concept-agnostic orientation because concept redundancy may be unavoidable in some secondary use contexts, so we use "concept" and "term" somewhat interchangeably.

Concept orientation is essential for vocabularies used in the capture of clinical data. It would be absurd, for example, to make care provider choose between ICD9 and ICD10 concepts in documenting patient conditions. Besides confusing the data capture process, it would compromise interpretation: e.g., choice between similar concepts appearing in both vocabularies might reflect a better match with the intended meaning, or it might reflect the provider's greater familiarity with one vocabulary. In the analytic context, however, it may be necessary to support overlapping, redundant, and even inconsistent vocabularies. A query over a data set including

records from before and after conversion from ICD9 to ICD10 might need value sets including codes from both. UMLS and OHDSI each provide a concept-oriented layer by which concepts and terms from any number of overlapping vocabularies are mapped to authoritative target concepts. But, according to our concept-agnostic orientation, this may not be necessary. OHDSI's vocabulary system, as mentioned above, accomplishes concept orientation by singling out certain concepts (or whole vocabularies) as "standard". But one might ignore this feature and see OHDSI's collection of vocabularies as an undifferentiated heap of concept-agnostic terms, leaving concept orientation as an exercise for value set designers and users.

While this might suggest a free-for-all, an abandonment of all hope for value set reuse, our aim is quite the opposite. With many vocabularies, many data sources, many different disciplines, industries, and use cases, the "same" concept will be representable with many different value sets. Some value set differences may reflect idiosyncrasies in regional or medical specialization coding practices, others will reflect actual nuances of meaning, and others still will reflect mistakes or oversights by designers. Our aim is to welcome differences in intended meaning or context-related code choice, while encouraging conformance, consolidation, and reuse whenever meanings are congruent and can be expressed appropriately for relevant contexts.

Ideally, provenance data of a value set can be captured in a standardized way to represent its intended meaning or context information. Machine learning algorithms may also aid in construction, consolidation, curation, retrieval, or evaluation of shared value sets, but human researchers and analysts must ultimately judge whether a value set fits their intended concept and

context. An interface for value set management, according to this principle of concept-agnosticism, would assume the role of facilitator, not arbiter, in determining concept congruence.

2.5 Standards, Infrastructure, and Design Recommendations

The following recommendations are intended to support the development of platforms that more effectively support reuse of semantic and analytic resources. While not comprehensive, they serve as a starting point for a more detailed and thorough set of guidelines to make reuse the norm rather than an easily ignored technical affordance.

Value set specifications and functional requirements. HL7 is currently balloting[6] a specification that identifies a standardized approach to value set metadata and structure: *Characteristics of a Formal Value Set Definition, Release 1* [83]. This specification has been the basis for HL7's Fast Healthcare Interoperability Resources (FHIR) value set resource. OHDSI's requirements are not represented in relevant HL7 working groups, and the OHDSI Architecture working group is not considering external standards in its value set specification development process. Even if HL7's specification is too detailed and complex to helpfully inform OHDSI's specifications, an important opportunity will be lost if no effort is made to compare value set specifications across these organizations and domains and explore possibilities for shared standards.

Definition processing and resolution. Value set definitions are taken as rules that must be applied at "runtime" in the context of a specific vocabulary collection, at which point they are resolved to a list of codes actually occurring in that vocabulary collection. There are multiple

[6] At the time of writing/publication, 2017–2018.

approaches to defining value sets: *by enumeration* of codes selected by an analyst or copied from an external source like a published study; *by rule*, e.g., a SNOMED CT code for angioedema and all its descendants; b*y composition* including set operations (union, intersection, difference, complement) or modifications of existing value sets. A single value set definition may refer to *multiple vocabularies*, and a resolved value set expansion may include codes from multiple vocabularies.

Standardized metadata. A value set requires more than an executable definition. Metadata standards should include: value set name, vocabularies referenced, vocabulary versions required if any, description, comments, links to external sources (e.g., citations for publications, URLs for value sets copied from online repositories), links to public use of value set (e.g., eCQMs), and *provenance tracking* of author information, dates of creation and modification, detailed documentation of successive user actions involved in crafting definition, readable presentation of ancestor provenance, as well as documentation of user attempts—successful or not—to locate appropriate value sets to derive from.

Computably traceable pedigree should be enabled by storing references to the "parents" of value sets constructed by modifying or performing set operations on existing value sets. Parent value sets may themselves have been derived from earlier value sets, forming ancestry paths back to value sets that were created anew. These paths can be used for *composite definition processing* allowing value set definitions to be assembled and resolved by starting at the start of its ancestry path and successively applying changes or set operations at each step, as well as for *provenance documentation*. For various reasons, the designer of a value set may want to make reference to other value sets for provenance documentation but not for definition processing.

Infrastructure and adoption. Real-world reuse will depend on adoption of software platforms and value set repositories supporting common specifications.

License-compliant openness. Value sets are composed of codes from controlled vocabularies, many under restrictive licenses. VSAC requires a UMLS license and user authentication for access to any value set. OHDSI authenticates licensing only for restricted vocabularies. A maximally open but legal reuse platform would accommodate vocabulary collections customized to users' needs and permissions, perhaps redacting license-protected codes from as necessary.

Open, public, crowdsourced curation. Where redundant value sets cover the same concept, they might be merged or one may be favored over others (in value set repository searches) based on evidence of being more widely used or preferred, e.g., by authorities recognized by user configuration. A process that provides shared, open value set definition will lead to improved vetting of the content and thereby ease the use of value sets not under an organization's direct control.

Network effects. To state the obvious, if there were already a platform and collection of value sets that everybody used or contributed to any time they needed a value set, that would be a powerful incentive for reuse. Conversely, even a perfect platform with every desirable affordance for reuse will face an uphill struggle until adoption reaches critical mass. The point here is that the allegiance of a user community can be as valuable in itself as any technical affordance, and these recommendations should not be taken as encouragement to build brand new platforms, but as a point of reference to facilitate efforts to *engage existing communities with value set platforms and repositories*, including, perhaps, commercial vendors as well as the non-commercial efforts we've brought up. Even if the existence of multiple platforms or

repositories is inevitable or necessary, *opportunities for synergistic cooperation on harmonization or consolidation projects should be sought and encouraged*.

Open standards, resources, and governance. Because of the power of network effects, communities may vie for control of standards, software repositories, or curation of value sets and other shared resources and repositories. Jaron Lanier [84] describes how companies scramble for the winner-take-all spoils of controlling "siren servers", central hubs for the sharing of crowd-sourced data. Technology supporting decentralized resource management may be needed to gain trust and participation.

Interactive, information-rich, high-performance visual interfaces. Given the range of formidable social and technical challenges facing value set reuse, especially regarding the ease of constructing one-off value sets, a successful platform will need interface innovation that goes beyond minimizing the cognitive and logistical costs involved in sharing and provides immediate positive benefits to users.

Modular components for integration into health analytics development environments and other analytic interfaces. Value sets are not ends in themselves; they are the computable representation of clinical concepts needed for other analytic tasks. An interface for creating, retrieving, using, or modifying value sets should be embedded unobtrusively into the context where value sets are needed. Users should see how their value set selection or modification choices affect the analytic task at hand immediately if possible.

Semantic graph visualization linked to local patient data. Designing an interface for semantic exploration, understanding, and navigation is challenging with some individual vocabularies (e.g., ontologies like SNOMED CT), and more challenging with a large collection of

vocabularies with intra- and inter-vocabulary hierarchies and mappings. An interface should allow the user to: efficiently, intuitively, and flexibly display the semantic neighborhood surrounding a set of codes; efficiently, intuitively, and flexibly display observational data matching currently selected codes; visually compare similar value sets (e.g., the current value set and the same after some modification), in terms of both semantic neighborhood and matched observational data; receive computer-aided simplification prompts, e.g., if a subset of codes can be represented by including some single code and all its descendants (or relatives by some other relationship like mapping or indication), that substitution should be recommended to the user; view and explore provenance execution plan and derivation tree documentation; receive prompts to examine and make use of existing value sets matching or similar to the one being designed.

2.6 Limitations

The perspective on semantic interoperability of value sets presented here and the design ideas reflected in our recommendations have been shaped by our work as academics and professionals. While a systematic survey and wider use case analysis, literature review, or environmental scan might have resulted in a better representation of the informatics community at large, the insights offered here are informed by our long and diverse experience working on these issues.

Our presentation of practices surrounding secondary use of health data is lopsided; most significantly by ignoring all but coded data. The development of reusable analytics for handling laboratory results, for instance, presents problems not touched on here.

Though many of the observations and ideas presented here were formed in the course of professional work (much of it for organizations in the OHDSI community), the paper has been written without funding or specific institutional sponsorship. This is reflected in our focus on

non-commercial efforts, CDMs, and OHDSI in particular. Our preference for open access standards and open-source software should also be noted.

Chapter 3.
Code Sets, Value Sets, and Phenotypes, Oh My![7]

3.1 Introduction

The growth of distributed research networks (DRN) — such as The National Patient-Centered Clinical Research Network (PCORnet) [14], Observational Health Data Sciences and Informatics (OHDSI) [85,86], Sentinel [13], Mental Health Research Network (MHRN) [87], Accrual to Clinical Trials (ACT) [88], and electronic Medical Records and Genomics (eMERGE) [89,90] — has increased the speed, frequency, and quality of research investigations using multi-site electronic health record (EHR) data. Analysis of repositories of coded patient data involves data models, coding systems, analytic protocols and implementations of the study's query logic. With the development of DRNs has come an explosive growth in shareable tools and resources—which have brought new capabilities and efficiencies beyond DRNs, to single-site studies, and to analyses for purposes other than research (e.g., clinical quality or value-based care, clinical guideline compliance). New tools and resources have required new language to describe them, leading, at times, to confusion, duplicate work, or miscommunication when collaborating or sharing resources for health data analysis.

We believe that a statement of definitions would benefit all the communities where research, public health, and administrative or economic analysis are performed on patient data collected

[7] Written in 2020 by Sigfried Gold, Luke Rasmussen, Laura K. Wiley, and Rachel Richesson but never submitted. Between 2019 and 2023, I favored the term code set over value set and felt there were important distinctions to be made between them. I switched back to using value set as a default only because some important people in my professional and academic circles have convinced me that the difficulties I had in publishing the content in Chapter 4 and Chapter 5 were due to the unfamiliarity of "code set" to many in the field. I include this glossary paper here because 1) it can act as a reference and brief synopsis of concepts used throughout the dissertation; and 2) because the distinctions laid out here are used in Chapter 4. (And were used in Chapter 5 until I switched the text to use "value set."

through the routine provision of clinical care [91]. In this article, we attempt to disambiguate and clarify a few essential terms in the form of a citable glossary that can be used across research, regulatory, and health administration communities.

Our glossary (summarized in Table 2 and detailed in the remaining text) specifically addresses the analysis of patient data coded using standard medical vocabularies. For each term in the glossary, we provide common acronyms, synonyms and definitions. Additionally for some terms, we identify challenges that we have observed in their use and propose potential solutions in the form of qualifiers that can be used with common terms to clarify their meaning, or alternative terms that have a broader scope than those commonly used in the literature.

3.2 Glossary

Table 2. Glossary of terms regarding electronic health record (EHR) data used within distributed research networks (DRNs)

Term (Acronym)—Synonyms *Definition*	**Challenges with Current Usage** **Proposed Solutions**
Real-World Data (RWD)—secondary use, observational, patient data. *"[D]ata relating to patient health status or the delivery of health care routinely collected from a variety of sources, such as EHR and administrative data [92]."*	Overly broad when speaking specifically of applications with code sets, given the multiple modalities RWD is collected in (structured data, text notes, images). Qualify with data provenance when scope is more specific: **structured data, coded data, coded with standardized vocabularies.** For non-coded data, meaning will usually be made clear anyway (e.g., unstructured, narrative text, device readings, image data).
Real-World Evidence (RWE) *"Clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD [92]."*	Overly narrow in some contexts, as it describes the resulting discovery (evidence), but does not capture the process by which the evidence is derived. Though analytic uses of RWD include of public health surveillance, clinical quality measure development, ad hoc queries of a clinical data warehouse, administrative reporting, RWE is not generally used that broadly. Use **analysis** instead if appropriate.
Analysis—real-world evidence generation, observational study, public health surveillance, clinical quality measure, query, administrative reporting. *Formal and informal efforts to answer questions based on RWD.*	
Computable phenotype—phenotype algorithm, electronic phenotype, cohort, patient population, EHR condition definition, EHR-based phenotype definition, phenotype, comorbidity or variable definition. *A computable phenotype is an algorithmic representation of a "measurable biological (...), behavioral (...), or cognitive markers that are found more often in individuals with a disease or condition than in the general population ... that can be determined solely from the data in EHRs and ancillary data sources and does not require chart review or interpretation by a clinician [93]."*	Highly overloaded with synonyms that are used inconsistently across fields. Use **patient group algorithm** instead if appropriate.
Patient group algorithm—patient group specification, patient group. *A broader term for computable phenotype, referring to any component of an analysis which attempts to associate a named real-world phenomenon with an algorithm for identifying it in RWD.*	

Value set—clinical code set, concept set, code list, term list, enumeration. *"In the context of terminologies and coding schemes, a value set is an uniquely identifiable set of valid values that can be resolved at a given point in time to an exact set (collection) of codes [94]."* *"[A]n unordered group of distinct clinical codes taken from a clinical terminology [95]."*	Clarifying the ambiguity of this term is the primary purpose of this article. Qualify with **permissible** when referring to value sets or code sets being used to constrain permissible values for data elements during data entry. Qualify with **analytic** or **phenotypic** when referring to analytic use with RWD.

3.2.1 Real-World Data

This is the data from EHRs and claims, as well as patient-reported outcomes and study questionnaires. Given the pragmatic nature of RWD we recognize that it is collected in heterogeneous modalities (with purpose), although our focus is on data coded using standardized, controlled, medical vocabularies. RWD is grounded as a preferred term by the 2018 JAMA Viewpoint article, "Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness [92]."

RWD, as a term, can be used to distinguish clinical data collected in practice from data collected in prospective studies or randomized clinical trials (RCT); but it is increasingly common (e.g., in pragmatic trials [96]) for RWD and RCT data to be combined.

3.2.2 Coded, Controlled, and Standardized RWD or Structured Data

RWD and computable phenotypes may reference EHR data that is not expressed in codes from standard clinical vocabularies; they can, in fact, be composed entirely of non-coded data. In discussing value sets, we are necessarily referring to RWD and phenotype algorithms that consist at least partly of coded data, and anything asserted about value sets will not be applicable to the non-coded data parts.

Adding the qualifier “coded” to RWD, and making clear in discussion of RWE, other analyses, and phenotypes, that they are being applied to coded data, sometimes makes a crucial distinction.

If the codes used in a particular analysis are not from a controlled or standardized vocabulary, it should either be clear from context or ideally stated explicitly. One may object that coded RWD should not be assumed to use only codes from controlled and standardized vocabularies, but in practice we believe such an assumption is reasonable in the absence of any indication to the contrary.

3.2.3 Analysis (of patient, health, or real-world data)

While RWD is a perfectly serviceable term beyond its originating drug safety and research context, the associated term, Real-World Evidence (RWE), seems a poor fit for some non-research analytic uses of RWD. RWE can encompass comparative effectiveness, patient-reported outcomes, and other sorts of observational studies, adverse event detection or verification algorithms, syndromic surveillance or pharmacovigilance. While not as felicitous as RWE, the term analysis is a readily understood, non-jargon term, and can comfortably cover clinical quality measurement, patient safety monitoring, cost-benefit or value-based care reporting, and ad hoc population queries by care providers seeking evidence to guide treatment decisions. For any analysis aimed at publication or performed using established research methods or study design, RWE is appropriate and might be preferred, but when the intended meaning is broader, RWD Analysis may best serve as a blanket term, applicable equally to research and operational uses of RWD.

The types of analysis mentioned above all require some definition of a patient population under study. This is usually done, at least in part, by selecting lists of codes from available data elements in the EHR. A number of terms exist for this process, each with its own technical definition and in many cases different nuances of meaning depending upon who is using the term.

3.2.4 Computable Phenotype

The term phenotype arises from biology where it denotes the observable traits of a living organism, arising through the interaction of genotype (the genetic blueprint or predisposition) and the environment. The biological phenotype refers to the expression of genotypes in the world—the result of nature and nurture. In the realm of clinical research informatics, the computable phenotype is the *digital footprint* of the biological phenotype. A phenotype can be "understood as measurable biological (physiological, biochemical, and anatomical features), behavioral (psychometric pattern), or cognitive markers that are found more often in individuals with a disease or condition than in the general population [93]."

Various phenotype terms have emerged to describe phenotypes as they are discernable in data, usually by adding a prefix of **electronic**, or **e-**, but often no prefix is used at all. They are sometimes called **phenotype algorithms**, linking it to the underlying query specifications or software implementation.

We consider the computable phenotype to be an electronic representation that (given our choice of term) is interpretable by a computer. It may be manifested in a CDM-specific analysis platform (ATLAS [97] for OHDSI, or i2b2 [98,99] for ACT), a transportable format such as the

Clinical Quality Language [100], or as a specific implementation in a given database system or programming language.

We do recognize that many phenotype algorithm definitions are represented in a non-computable format, such as descriptive pseudocode available on PheKB [101]. We are purposefully focusing on computable definitions here to highlight the importance of value sets in their definition.

In the context of RWD analysis, **phenotyping** is the gerund or present participle of phenotype, referring to the practice of constructing or working with computable phenotypes.

3.2.5 Patient Group Algorithm

Some researchers use "phenotype" to refer to a wider range of phenomena than those that would be called phenotypes in a biological context such as exposure to a specific drug or class of drugs, patient transfer from the emergency room to an operating room, patient membership in particular health plans. We offer no opinion as to whether the broader or narrower definition is correct, but we do offer "patient group algorithm" as an alternative term for the broader definition. It is not in common use, to our knowledge, but it is somewhat self-explanatory and will not conflict with the narrower definition of computable phenotype. For example, in an observational comparative effectiveness study, separate patient groups would be specified for each required cohort, treatment, comparator, outcome, or other study element, as well as for factors used to stratify patients within cohorts or, perhaps, construct propensity scores.

The colloquial use of the term "Phenotyping", as in "members of the phenotyping community," should remain serviceable, as the practice of phenotyping will generally require

some patient group algorithms that wouldn't be considered phenotypes by the narrower definition.

3.2.6 Value Set

"Value set" is challenging both because it has contrasting meanings that sometimes need to be distinguished, and alternative terms that are sometimes preferred, such as concept set or code list.

After defining value set (see table above), Pathak et al., in their 2008 LexValueSets paper, go on to say,

> Value sets exist to provide possible values for data elements (i.e., variables) that can be present, for example, in an electronic health record. The main objective of modeling value sets is to specify a concept domain with certain slots or attributes of interest such that the attribute-values can be obtained from one or more terminologies of interest [94].

The use to which they put value sets in this paper is not phenotyping but the sharing and re-use of "value sets across multiple medical information systems."

The Williams et al. use case for the 2017 "Clinical code set engineering for reusing EHR data for research: A review," on the other hand, is clearly for phenotyping: clinical code sets are

> an unordered group of distinct clinical codes taken from a clinical terminology...[that] typically relate to a specific medical concept and these codes would be used by researchers to construct queries to execute against a database to extract patients or data for use in further analysis [95].

Williams explicitly recommends using "clinical code set" or just code set in order to distinguish code sets based on standardized clinical vocabularies from others.

Gold et al. went to some pains in their 2018 paper, "Clinical Concept Value Sets and Interoperability in Health Data Analytics," to be clear without unambiguous terms at hand:

> The phrase *value set* is problematic. Our usage may be confusing to those familiar with value sets as criteria for populating drop down lists or for constraining the values allowed in a data element. The term may also be unfamiliar to health data researchers and analysts who routinely construct value sets to query encoded data but call them by a different name (e.g, code lists or concept sets) or may not recognize them as distinct components of analytic algorithms at all. [51]

We propose these qualifiers: **permissible** for lists of codes that constrain the values allowed in a data element, and **analytic** or **phenotypic** for lists of codes used to extract similar patient records from a clinical data warehouse for use in a computable phenotype or analysis. The distinction may not be important for some researchers, but we have encountered significant confusion when speaking of phenotypic code sets with interlocutors who thought we were talking about permissible value sets. In respect for Williams' recommendation, we try to use code set in analytic contexts, and "value set" when speaking of clinical terminology services or permissible values; but we also consider the conventions of our audience. Using one of these prefixes should clarify one's meaning whichever of these terms is used.

3.3 Conclusion

A final distinction we would like to make within these terms is specifically around the relationship between computable phenotypes and analytic code sets. Phenotypes that refer to coded data will include at least one analytic code set, generally along with other logic, but a plain code set by itself can be used as a phenotype. It is important to clarify this relationship, particularly with respect to validation: a code set validated for use as a phenotype by itself (or as a permissible value set) should not be assumed for use as part of a more complex phenotype without additional validation in that context.

The definitions we have presented are meant to be descriptive, not prescriptive. We cannot dictate *how* the various components and processes of health data analysis are performed or the

terms to denote them, recognizing that there is often a rationale and history for preferred terms. Still, we hope publication of this glossary will save authors facing the usage challenges described here from having to start from scratch in solving them.

Chapter 4. Real-world analysis of the analysis of real-world data: A field study[8]

4.1 Introduction

As informaticists, epidemiologists, clinical researchers (not to mention regulators, health system administrators, software developers, and health economists), we believe patient data in clinical data warehouses (CDW) and distributed research networks (DRN), if properly analyzed, can yield transformational insights about patient populations and individual patients, and advance our knowledge of the biological, social, and institutional phenomena that affect patient health. Without delving into questions of data capture or the reliability of real-world data (RWD) or routinely collected data analysis generally, this paper addresses the construction, validation, and reuse of code sets, one of the fundamental building blocks of RWD analyses [2].

One might argue that the electronic phenotypes, the algorithmic expression of a clinical phenomenon of interest for an analytic task, are *the* fundamental building block of RWD analysis. Code sets are generally the primary components of electronic phenotypes. The central focus of this paper is not the phenotype but the code set[9], which often serves as a phenotype's primary building block. We recognize that coded data—data captured in the form of concept identifiers from controlled medical vocabularies—by themselves are seldom sufficient for

[8] This chapter combines material from three different sources, all focused on a general presentation of my field study of value set development practices (UMD IRB #1405794-8.) It includes parts of my integrative paper (the UMD iSchool's requirement for advancing to candidacy, like a qualifying exam), my dissertation proposal, and a paper submitted to The Journal of Biomedical Informatics and preprinted on medRxiv [20].

[9] In the previous and subsequent chapters, the term "value set" is used regardless of context, sometimes distinguishing between permissible and analytic value sets for data capture and RWD analysis respectively. In most of this chapter, "code set" is used as the more general term (as recommended by Williams [17]) and is contrasted with "value set," which is used for specific contexts.

accurately identifying clinical conditions in RWD [102,103]. Many phenotype algorithms incorporate uncoded data such as narrative notes and numeric observations and involve conditional and temporal logic. Nevertheless, vocabulary codes are generally the starting point for analysts designing phenotypes.

With RWD, data capture is not controlled by the researcher, and vocabulary codes are not always used as intended or expected. Each code represents a term or concept from a controlled medical vocabulary. The collection of terms in the code set are *functionally synonymous*—meaning that querying records that contain any one of these codes should ideally match all and only the patients who have experienced the phenomenon of interest [18]. In order to approach this ideal goal, the code set may need to be adapted to idiosyncrasies in the patient data. Therefore, purely *semantic* validation of a code set by experts in the terminologies used and the clinical subject matter may not be considered adequate; *empirical validation* of patient data may also be needed.

Discussions of code sets and their validation and reuse in RWD applications have, we suggest, been muddled in prior literature because they do not distinguish the RWD context from applications where some control over data capture is possible, such as common data element (CDE) design and clinical quality measures (CQM), in which purely semantic validation may be sufficient. Empirical synonymy is not the same as semantic synonymy, so the validity of a code set in a phenotyping application depends on the data being analyzed, the patient population, local coding practices, clinical workflows, and more.

Concerns regarding the quality of code sets and the challenges faced by code set developers in identifying and selecting the set of codes best fitted to their clinical intention and analysis goals

have inspired the development of code set authoring and sharing platforms [28–30] and other scholarly and practical efforts to champion and facilitate code set reuse [15,16,35–38,31,39,40,25,27,32,104,42,41,51].

A 2017 paper by Williams, et al. [17] discusses problems with code set engineering that use RWD and the potential gains for study rigor and reusability if those problems are addressed. The authors provide nomenclature, a consolidated articulation of published knowledge on code sets, and a valuable catalog of recommendations for advancing technology for managing code sets.

Our study builds on that paper's insights to provide a process model that more accurately reflects the diversity of real-world code set development practices, and accounts for the specific contextual factors that shape the development process in different circumstances. To characterize the practice of code set development, we conducted an online survey, semi-structured interviews with a subset of survey participants, and observation where possible of participants building code sets.

Beyond advancing academic understanding of code set development, the contributions here offer practical aid in managing that process—accounting for the array of methods and practices available, their appropriate use in a given context, and how to systematically execute them. Further, it exposes opportunities for innovation in software to support recommendations made in prior literature.

In papers discussing code sets where the focus *is* clearly on the analysis of RWD, a single type of analysis often seems to be assumed: research for publication. This paper attempts to consider analytic tasks more generally—e.g., the support of clinical care, pre-research

exploration, health administration, patient safety reporting, pharmacovigilance, infectious disease tracking, and other forms of epidemiology or public health surveillance.

In order to understand the real-world practices surrounding the use of code sets, we designed and conducted a field study of epidemiologists, clinical researchers, or informaticists and others who work with RWD. We conducted an online survey, semi-structured interviews with a subset of survey participants, and observation where possible of participants performing actual work to create code sets.

Our study design intended to characterize the practice of code set development: the variety of institutional settings and analytic goals of code set developers, the data and software available to them, how contextual factors affected code set development and use. We were attuned to whether the concerns and processes shaping code set development were as heterogeneous as the contexts.

While there appears to be universal agreement that crafting high-quality code sets is difficult, our participants held very different ideas about how this can be achieved and validated. A primary divide exists between those who rely on empirical techniques using patient-level data and those who only rely on expertise and semantic data. Code set development practices relying on patient data must grapple with problems around private health information (PHI), software configuration, and questions of cross-database validity. Rigorous validation may require PHI, but the privacy and specificity of patient databases constrain possibilities for reuse. These tensions, we propose, have stymied past efforts to promote meaningful, widespread reuse of code sets for phenotyping.

Williams [17] breaks code set management down into four phases: construction, validation, sharing, and reuse. These terms are intuitively descriptive, and we continue to rely on them, but the practices found in our study could not easily be classified into one of these phases and necessitated the formulation of a structured process model, which is able to account for observed variability in formality, thoroughness, resources, and techniques.

Between the requirements planning and naming of an initially empty code set and the decision that it is complete, the diverse array of tasks involved in assembling and refining it comprise, according to our model (see Section 4.6), an iterative cycle of (1) code collection, (2) code evaluation, and (3) code set evaluation. When this cycle ends, the code set is (4) accepted and ready for use. These steps are sometimes followed by (5) reporting—for publication, documentation, or stakeholder requirements—on the methods and results of creating and validating the code set, and maintenance as future vocabulary versions are released.

We analyze the real-world practices and ideas of code set developers and synthesize these with accounts of code set validation and reuse in the literature, as well as our own professional experience. Our findings problematize understandings of code set validation and reuse while offering novel perspectives and approaches to capturing and sharing evidence to support trust in the quality and fitness for use of code sets and metadata being considered for reuse.

4.2 Background

4.2.1 Value Sets

A list of codes representing the variety of terms used to represent a clinical condition or event included in an analysis or study (e.g., an outcome, confounder, treatment) is known as a *code set* or *value set*.

The codes in question are from controlled, standardized medical vocabularies and represent diagnostic conditions, medical procedures, drugs, and other clinical concepts or terms itemized in those vocabularies. These codes are entered into EHRs or administrative claims databases and are then used in research (and other analytic projects) to query those databases for groups of patient records.

Much of the value set literature comes from scholars and practitioners engaged with terminologies and ontologies, terminology standards, and terminology services—some of whom participated in the burst of foundational study of medical vocabularies in the 1990s and 2000s, e.g., [46,47,67].

The early 2010s saw a flurry of papers centered around problems with value set quality and offered methods for evaluating them [35,36] and efforts to encourage their sharing, reuse, and standardization [28,29,105]. This work mostly addressed the use of value sets in defining clinical quality measures (CQM) or specifying the domains for data elements used in clinical trials.

With the importance of value sets gradually being recognized by the clinical research community, standardization of value sets is becoming imperative, as it can enable comparison across disparate datasets and facilitate reuse of well-defined value sets to advance clinical research studies [43].

An explicit interest in RWD is not a feature of these papers, with the exception of the UK paper [30] quoted above. Around the same time that the NLM was releasing VSAC primarily to house value sets contained in the 2014 Meaningful Use criteria [36], Clinical Codes in the UK

was launching a repository to share code sets used in published electronic medical records studies.

The American terminology experts working on value sets undoubtedly had RWD studies in mind as an important use case, but they may have considered these sources too idiosyncratic and their quality insufficient to benefit from shared value sets unless such value sets were also used to constrain permissible values at the point of data capture. The following criteria, for instance, seems to assume a high degree of standardization and semantic coordination of terminologies, data models, value sets, and data elements:

> A value set should contain all the relevant codes for a particular data element. Moreover, the value set name should also denote this data element. From a *terminological perspective*, the code corresponding to the data element in the code system should be present in the value set, along with all its descendants. As a consequence, the value set is expected to be rooted by one concept and to contain all the descendants of this root concept. [36] [emphasis ours]

This quote recommends that every value set correspond to the permissible values for a particular data element and that it consist of a single code and its descendants. This may or may not be reasonable as an expectation for value sets generally, but it elegantly encapsulates the goals and assumptions of the value set literature—which might be better labeled the *terminological perspective*.

4.2.1.1 Extensional and intensional definitions and expansion

As standards organization Health Level Seven (HL7) defines it, a value set *definition* is "a description of the set of Concept Representations (usually codes) that are intended for use," and its *expansion* is "the set of resulting codes actually obtained for a particular use, drawn from one or more specific Code System versions." The definition can take one of two forms, *extensional*:

"explicitly enumerating each of the Value Set concepts," and *intensional*:[10] "defining an algorithm that, when executed by a machine (or interpreted by a human being), yields the desired set of elements [25 p. 16]."

An intensional definition can include rules like:

- a code plus all its descendants,
- the descendants of a code but not the code itself,
- the exclusion of a code and its descendants (from the codes resulting from other rules), or
- the codes in one code system (e.g., SNOMED-CT) that are mapped-to from a code in another vocabulary (e.g., ICD10).

4.2.1.2 Metadata

The selection of metadata stored with a code set, if any, generally depends on what is allowed or required by the repository or data format used.[11] The FHIR value set resource definition [106] allows for an extensive collection of metadata, while the OHDSI/ATLAS concept set tab allows for little more than the concept set's name. Metadata [26] that may be captured can include information about:

- Intention (e.g., clinical meaning, purpose, or context of use);
- Expansion (applying intensional rules of a code set version against a code system version to get resulting codes);
- Provenance (e.g., authors, stakeholders, dates, sources of content);
- Evaluation or validation (description of any validation tests or evaluation activities performed and results).

[10] Intension (with an s) is a term of art. We spell intention with a t when referring to the intentions of humans and intension with an s when referring to a rule-based method of defining a code set.

[11] See Alper, et al., 2022 [26] for a more thorough discussion of metadata for code sets and other computable biomedical knowledge artifacts.

4.2.2 Computable phenotype definitions

To the degree that established guidelines for the conduct of observational studies based on RWD exist, the RECORD Statement, along with STROBE [107], which it extends, is a foundational document. Prominent amongst its directions are:

> **RECORD ITEM 6.1:** The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided.
>
> **RECORD ITEM 7.1:** A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these codes or algorithms cannot be reported, an explanation should be provided [5].

What the RECORD Statement calls algorithms can go by various names, including: **computable phenotype definitions**, cohort definitions, or just patient groups. For some authors, "phenotype" is used more narrowly to refer to disease conditions and not, for instance, exposures. We use it to cover the operationalization of any set of clinical criteria for patient record selection [19]. (See Chapter 3.)

Richesson [15] also draws a helpful and clarifying distinction between phenotypes, computable phenotypes, and computable phenotype definitions, though phenotype by itself is often used loosely to refer to any or all of the three.

Figure 1, copied from that paper, shows a table of features from a comparison of seven phenotypes from different sources, all for type II diabetes mellitus. The eight "Data domain criteria" columns represent the different features. Rows containing only a green circle represent phenotypes consisting of nothing but a single code set in a single domain. Rows containing triangles represent more complex phenotypes that include multiple code sets and additional logic.

Table 1 Data domain criteria used in selected phenotype definitions

Phenotype definitions:	Data domain criteria: ICD-9-CM 250.xx	ICD-9-CM 250.x0 and 250.x2 (excludes type 1 specific codes)	Expanded ICD-9-CM Codes (249.xx, 357.2, 362.0x, 366.41)	HbA1c	Fasting glucose	Random glucose	Abnormal OGTT	Diabetes-associated medications*
ICD-9-CM 250.xx	●							
CMS CCW	▲✳\\		▲✳\\					
NYC A1c Registry				●				
Diabetes-associated medications								●
DDC		▲	▲	▲\\	▲\\	▲\\	▲\\	▲
SUPREME-DM	▲✳\\		▲✳\\	▲\\	▲\\	▲\\	▲	▲
eMERGE†		●✳\\		▲	▲	▲		▲

*Medications vary by phenotype definition and are listed for each in the supplementary appendix (available online only).
†The eMERGE phenotype definition consists of five case scenarios with varying combinations of criteria. Any instance of type 1 specific codes (ie, 250.x1, 250.x3) results in the exclusion of the patient.
●=Sole criteria.
▲=Optional criteria, one of many.
✳=Distinction made between inpatient and outpatient context.
\\ = Distinction made for multiple instances and/or time points.
CMS CCW, Centers for Medicare and Medicaid Services Chronic Condition Data Warehouse; DDC, Durham Diabetes Coalition; eMERGE, electronic medical records and genomics; HbA1c, hemoglobin A1c; ICD-9-CM, International Classification of Disease, revision 9, clinical modification; NYC, New York City; OGTT, oral glucose tolerance test; SUPREME-DM, Surveillance, Prevention, and Management of Diabetes Mellitus.

Figure 1. Different diabetes phenotypes compared [14]

Phenotypes generally consist of algorithmic logic to select records representing patient characteristics or events, and generally include one or more code sets. They can take the form of SQL queries; R, Python, or SAS routines; queries constructed through specialized user interfaces; or any other way records might be processed in the attempt to match the researcher's clinical criteria.

A *phenotype algorithm* can be as simple as a single code set with no additional logic, in which case the presence or absence in the patient's medical records of one or more codes in the set determines the patient's inclusion in the phenotype. More commonly, phenotypes are more complex than that.

Figure 2 shows a more complex phenotype from eMERGE, a "national network organized and funded by the National Human Genome Research Institute (NHGRI) that combines DNA biorepositories with electronic medical record (EMR) systems for large scale, high-throughput genetic research in support of implementing genomic medicine." This diagram of a sample eMERGE Network phenotype shows, in yellow, a set of inclusion criteria and two sets of exclusion criteria for a phenotype for finding European-ancestry patients with autoimmune hypothyroidism. Specific criteria from each of the three sets appear in the green boxes, which mostly consist of code sets. For code sets drawing on the ICD9 or CPT vocabulary, the actual codes are given. For medications and lab tests, lists of medication or ingredient names or specific lab values would be converted to value sets when the phenotype was implemented. Those codes could not be practically listed in the figure as a single ingredient might occur in dozens of medications and thousands of distinct drug products [108].

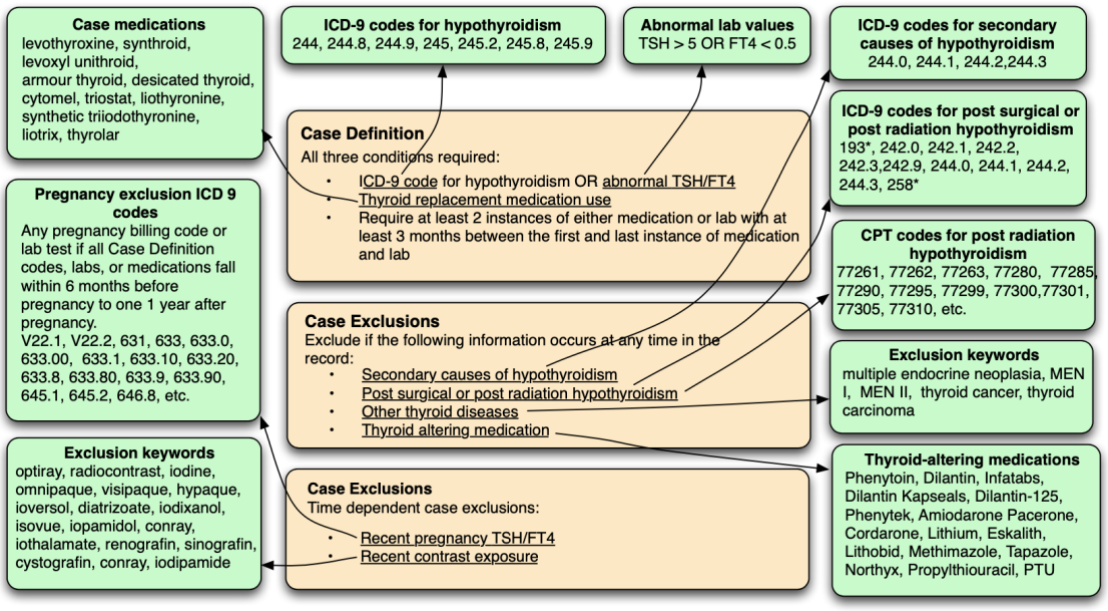

Figure 2. Diagram of an eMERGE Network phenotype [94]

4.2.3 Real-world evidence, phenotypes, and code sets

It is possible to implement a simple, ad hoc study in one or two SQL queries, which might blur the line between phenotype and study algorithms. Even in a complex, structured system like OHDSI where cohort definitions are independently defined objects that are given as parameters to study algorithms, it may be a matter of judgment whether a given rule should be part of the study algorithm or included in the cohort definition.

So, study algorithms and phenotype algorithms *can* be conflated, but this leads to the confusing and misleading idea that code sets, by themselves, represent "clinical concepts". It would be nice if it were so (and when code sets for analysis are equivalent to permissible value sets for data capture, it can be so), but it has been repeatedly shown that simple diagnostic code sets perform poorly as cohort definitions compared with multi-modal algorithms [102].

4.2.4 Permissible value sets and analytic code sets

"Clinical code set", "code set", "concept set", and "value set" are generally used synonymously (see [17, pp. 1, 2, 9]). Different participants in our study preferred one term or another but appeared to use them interchangeably. The focus of our study on the use of code sets in the analysis of real-world data was made clear to our participants in the survey text and during interviews. Nevertheless, many made no distinction between these and permissible value sets, i.e., lists of values (codes) constraining the permissible contents of a data element. We have been unable to find the distinction between permissible value sets and analytic code sets articulated in the literature, though their conflation seems to cause confusion, especially around validation.

A permissible value set constrains a data element or domain to the universe of semantically distinct terms it can contain. In contrast, an analytic code set is meant to hold a set of functionally equivalent terms, any of which indicates an instance of a more general concept.

Value sets for CQMs are a special case where the terms are functionally equivalent, but the CQM authority dictates which terms will be accepted. Authors may sometimes consult EHR data in quality measure design, but these measures tend to be applied regionally or nationally, not to any specific database, and it is the responsibility of the institution running the CQM to assure that data are coded to match the CQM's value sets. For that reason, we classify these as permissible value sets and CQMs as non-RWD applications.

As a contrived illustration, a permissible value set for gender might be: Female, Male, Other, Refused to answer, Unknown. Each term is distinct. An analytic code set for female might be (assuming case-insensitivity): female, f, fem; all synonymous and interchangeable for query purposes. And an analytic code set for non-male might be: female, f, fem, other, non-binary. Clearly, f, other, and non-binary are not synonyms, but in reference to the concept of non-male, they are equivalent.

There are many ways to validate value sets in contexts outside RWD analysis, some of which we explored in the study but are out of scope for this paper. According to a 2012 paper, "Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups,"

> With the importance of value sets gradually being recognized by the clinical research community, standardization of value sets is becoming imperative, as it can enable comparison across disparate datasets and facilitate reuse of well-defined value sets to advance clinical research studies [43].

The importance of facilitating and eventually standardizing code set reuse, recognized in permissible values contexts, is also recognized by some in the RWD or phenotyping community [17,30]. It appeared to be taken for granted by many of our participants. Prior to conducting this study, we had recognized and described some of the obstacles to reuse [18,42] but had not yet grasped that reuse depends on communicable validation reporting.

According to one of our study participants (P16), "We don't see value sets discussed a lot in the phenotyping literature." This is true in part because the phenotyping literature often conflates code sets and phenotypes, referring to both as phenotypes. This is not wrong—as long as their purpose is to select patient records matching some clinical idea, a code set with no other logic can be used and validated as a phenotype. That is, the criterion for being a good one is that it selects the records the researcher would want it to select for the analysis they are doing and that it does not select other ones.

However, the value set literature has a different take on quality and validation. The difference between their terminological perspective and, what we might call the "empirical perspective" of the phenotyping literature, can be understood by drawing an analogy to prescriptive and descriptive grammars and lexicons: the terminologists give us the vocabularies, data elements, and value sets we *should* use, while the phenotypers try to give us algorithms that match the codes and patterns that *are* used. Though prescriptivism in general language has lost favor over the past century or so and tends to be seen more as enshrining the linguistic habits of an elite social class than providing a guide to more general usage. But in medical informatics, prescriptivism is completely necessary: without efforts to standardize medical language use, much of the potential of algorithmic processing of medical language would be lost.

By regulating the expression of medical observations through careful semantic constraints on the process of data capture, prescriptive approaches to terminology and circumvent the need for empirical validation entirely—assuming, that is, that the clinical domains defined for data capture correspond to those under study by the analysts.

We can distinguish five types of value set implicitly indicated across the various literatures, three belonging to the **terminological perspective**:

1. Purely semantic; designed by terminologists, ontologists (or perhaps EHR vendors) with clinical expertise to represent all the varieties or features of interest for a particular clinical concept or phenotype;
2. A permissible value set for a common data element or a data element used in a specific application or clinical trial;
3. A value set approved by an authority (e.g., the National Quality Forum or CMS) for use representing a particular clinical condition in one or more clinical quality measures;

And two belonging to the **empirical perspective**:

4. A code set-only phenotype;
5. A code set for use in a phenotype that contains other logic (and perhaps other code sets) as well.

We order them 1-5 because they form a kind of continuum of confusion. They are all value sets, so their digital structure need not differ at all, and they all *could* appear in a code set repository. Types 1 and 5 are clearly different sorts of objects, though a type II diabetes mellitus value set of the two types might share many or all of the same codes.

A type 1 or 2 value set, if it adhered to the ideal expressed in the [36] quote above, would consist of a single root code and its descendants. Or, if it used codes from vocabulary not designed with ontological discipline, it might consist of codes from different subtrees, but could have validity by virtue of the terminologists attesting to its completeness and accuracy.

A code set of type 5, however, could not meaningfully be validated at all without executing the additional logic of its containing phenotype algorithm, and in reference to databases where that algorithm has been validated.

Types 3 and 4 are more similar to each other: since each represents a clinical condition without additional logic; presence of any code from the set in a patient record indicates presence of the clinical condition, and absence of the code implies absence of the condition. Nevertheless, the empirical validation normatively expected of a phenotype algorithm is not possible or necessary with type 3. The authors of a type 3 value set might perform some validation with real-world data, but that's not what gives the value set its authority. Since the clinical quality measure is to be reported by clinical practices and health systems nationally, it would not be possible to validate it. But the process of issuing a clinical quality measure can include public comment periods and other checks; and once it has achieved approval, it has the force of CMS reimbursement or some other authority behind it.

Code set types 4 and 5 appear in the phenotype literature, though usually in the service of explaining the composition of phenotype algorithms, not as objects holding any inherent interest on their own—with the exception of a few papers that do treat code sets as worthy of focused interest even in the context of RWD analysis.

4.2.5 Code sets for phenotyping as a digital artifact type worthy of focused study

In addition to Springate [30], these exceptions include [109], [18], and, of particular interest, Williams et al., "Clinical code set engineering for reusing EHR data for research: A review." It was the first paper, to our knowledge, to attempt a general, systematic discussion of code sets

and their management [17]. Williams conducted a broad literature review and identified 30 methodological papers on code sets in order to

> review and compare methods and tools for managing (constructing, validating, sharing, and reusing) sets of clinical codes reported in the literature; and […] develop recommendations for the management of clinical code sets and for the design and implementation of clinical code set management tools.

Though we find this paper invaluable and a welcome confirmation of our own high estimation of the importance of code sets in RWD analysis, it takes both the value set and phenotyping literatures on their own terms without commenting on their disparate perspectives. In describing approaches and recommendations for validating code sets, they offer: physician input and review of code sets to prevent type I errors; use of vocabulary hierarchies and iterative search to prevent type II errors; sensitivity analysis as codes are added or removed; and as a more thorough approach, having clinical experts review patient records to create a gold standard or reference dataset to which code set results can be compared to generate statistics such as sensitivity and specificity.

Gold standards or reference data sets are the recommended norm in the RECORD Statement and other guidelines, but the quality of these is often criticized. Williams warns that the data necessary for creating a gold standard may be challenging if not impossible to obtain and that the process can be prohibitively time consuming. They describe challenges of two general sorts: with acquiring the data necessary to build the gold standard; and with the amount of time and expertise required to accurately classify patients as manifesting the phenotype or not.

Some challenges with data acquisition are institutional and legal: HIPAA, IRBs, and institutional hurdles greatly restrict the availability of RWD and the analytic uses it can put to. A brief overview of these issues appear in “Clinical Data: Sources and Types, Regulatory

Constraints, Applications" [110]. Some are technical and infrastructural, having to do with the efforts required of informatics staff at a researcher's institution making such data accessible to both gold standard creators and to RWD researchers designing and implementing their analytic algorithms. These challenges do not apply to permissible value sets contexts.

4.2.6 Research Questions

Williams sees code sets as worthy objects of study, but do not distinguish the code set validation process from the phenotype validation process. This is distinction is crucial: empirical validation must be performed at the phenotype level (except in code-set-only phenotypes where there is no difference between the phenotype and code set levels).

It is problematic, certainly, to separate the code set management process from the management of the analytic algorithms the code sets are used in—*but,* the obstacles to making widely usable improvements to available software for managing analytic algorithms discourage progress that could be made by focusing design efforts on code sets and the relatively accessible and standardized resources needed for building them. We are not claiming that empirical evaluation or validation is not necessary, only that it presents challenges to code set authors, software designers, and systems developers, and that the need for it should not inhibit the development of tools that can improve analysts' ability to create high quality code sets.

We have been motivated by a belief that there is a lot of untapped semantic knowledge in vocabularies, groupers (such as Clinical Classifications Software Refined (CCSR) [111,112]), value set repositories, and other sources that could be used in value set construction, evaluation, and reuse with better interactive visualization tools, and that such tools would have a high impact on improving the quality of phenotyping and RWD analysis generally.

The specific research questions that motivated our study are:

- What resources, practices, tools, and infrastructures are used in managing phenotypic code sets?
- Does a different picture emerge when we look across RWD analysis contexts and use cases—not just formal, observational research studies for publication but for exploratory, ad hoc, or other internal analyses?
- Do we see patterns in uses, motivations, challenges, even across this wide spectrum?
- How are vocabulary and other semantic resources used in electronic phenotyping and other analytic use of RWD?

4.3 Methods

4.3.1 Study design

The mixed-methods study began with an in-depth electronic survey investigating the participant's experiences and current practice with clinical code sets in electronic health record databases. We followed up with a semi-structured interview with a subset of the survey participants, on details of their experience and practice with code sets. A few of these interviews were conducted in person, with the rest via Internet videoconference. Where possible we observed interviewees demonstrating their tools and processes for developing code sets.

4.3.2 Professionals as participants

The purpose of this study was to collect data about our participants' past experiences and current practices in the use of phenotypic code sets in order to identify barriers, gaps, and pain points that enable designing better tools to support this practice. For this reason, we obviously needed to engage actual professionals with routine hands-on experience with this kind of data.

While we do not go so far as to claim that our evaluation methodology is a form of *contextual inquiry* [113,114] or other formal types of ethnographic research, we do follow general principles of ethnographic methods for human-computer interaction (HCI) in that the best results

are gained from observing participants working on their own data in their workplace rather than taking them to the laboratory with artificial data. Also, the strength of our findings rest on the triangulation achieved between the survey, interviews, and work observation.

4.3.3 Recruitment objectives

Although our goal was to study professionals performing a certain task in the course of their work, we did not set limits on the variety of our participants' professions, nor on the context or motivations for their performance. The task, developing a code set for analysis of coded patient data, is necessarily subordinate to a larger task, defined abstractly as analysis; and the object being developed is a component of a larger object, a computable phenotype definition, which is itself a component of the algorithm implemented to accomplish the analytic task.

These larger contexts, tasks, and objects can be almost infinitely variable. Letting them vary naturally without constraint meant that our target population was large, diverse, and challenging to locate in the population at large. At the same time, our definition of membership is precise and has a compelling logic in relation to our original design goals.

Reaching a statistically representative sample of the entire target population would be challenging for a much larger study than ours and was not necessary for making significant advancement towards our goals. Rather than strive to be comprehensive, we prioritized reaching a sufficiently broad and diverse array of participants to encounter and explore some aspects of their work that are more or less consistent and common across contexts and use cases as well as aspects that are variable or inconsistent.

Also, since our goal was ultimately to inform software design to improve the quality and efficiency of code set development, and the leading edge in technology for RWD analysis tools,

techniques, and infrastructure is advancing very fast, our ideal participants would be highly skilled, well informed, and working with newer technology. Within those parameters, we sought participant heterogeneity along the following lines:

- *Purpose:* Drug safety (adverse event surveillance), comparative effectiveness, syndromic surveillance and infectious disease monitoring, clinical quality measurement, patient safety, health economics;
- *Institution type:* Academia, government, non-profit service providers and associations, and industry—pharmaceutical, health care systems, and services or software consulting;
- *Profession or role:* Physician, professor, academic professional, manager, analyst student, awardee of government contracts and grants;
- *Disciplinary approach:* Medical informatics, epidemiology, computer science, statistics, economics;
- *Project scope:* Single-institution analyses, studies using large collections of vendor-supplied data, national or global distributed research network studies.

4.3.4 Survey as filter and guide

A primary function of the survey was to filter our study population to include only people whose work involves using code sets for in analyzing health records. Since "code set," "value set," and other terms can easily be misunderstood, we needed to address risks of both type I and type II errors in participant selection, which could compromise our ability to understand participant answers. That is, without taking particular care to ensure that participants belonged to our intended population, there was a high probability that they would interpret our questions in ways we did not intend and provide answers that we might misinterpret or fail to understand. Or, alternatively, participants might incorrectly self-select out even if they even if they did belong to our intended population but misinterpreted the intent of our questions.

Rather than expect survey respondents to read a long discursive essay explaining terms, we attempted to craft questions that would orient them to our nomenclature, while at the same time

collecting information that would help us assure that they met our criteria and realized it. The survey questions are listed in Appendix A. Survey questions

We designed the survey so that earlier questions would implicitly define terms used in later questions. We did this by asking questions to establish that their work involved:

1. Coded data,
2. EHR or claims data collected in the provision of routine care,
3. Analyses of patients cohorts defined in terms of clinical characteristics (e.g., diagnosed with hypertension, or taking antidepressants), and
4. Phenotypic code sets.

Automated exits were coded into the Qualtrics survey instrument at each point where responses could indicate that the respondent did not meet our criteria.

4.3.5 Protocol

Our survey instrument, designed and hosted with Qualtrics, consisted of five parts: (1) a consent form, (2) respondent information, (3) a first qualifying question to assure that respondent works with coded data, (4) questions about their work with coded data and a second qualifying question to assure they also use code sets, and (5) open-ended questions about their work with code sets.

To be eligible, participants were required to be at least 18 years of age and consider themselves to be regular users of clinical concept sets and electronic health records in a professional capacity. Note that participation in any part of the study was entirely voluntary and self-selected.

For a subset of the survey respondents, we followed up with a standard semi-structured interview protocol [115] (see Appendix A). Building from a standard script, the interviewer identified or tailored questions for each participant based on their survey responses, with the aim

of clarifying and elaborating on survey answers where the meaning or implications did not seem self-evident.

Where possible we used contextual interviewing over Zoom or Skype using screen sharing to observe the participant working on code set tasks.

We performed thematic analysis, beginning with inductive, open coding on interview transcripts, responses to open-ended questions, and also on interview notes when recordings were not available.

4.4 Results and discussion

4.4.1 Survey

4.4.1.1 Population sampling

We used an initial convenience sample, then snowballing to reach our target sample sizes. Our survey recruitment invitation was distributed to almost 150 professionals believed to have specific experience working with clinical code sets. These came from 77 email invites from Qualtrics (to the first author's professional contacts or individuals identified as contributors to VSAC). From this set, 25 surveys were submitted with 22 fully completed and usable.

In addition to the initial 77 invites from Qualtrics, an unknown number of people were reached using survey links that were posted to the OHDSI forum and additional email invitations. Of this second round, 24 surveys were submitted with 14 fully completed and usable, all of which included respondent-entered name and email.

From the 36 fully completed surveys, we conducted 15 one-hour to ninety-minute interviews. Eight of these interviews were recorded and transcribed. Detailed notes were taken for all 15 interviews.

We were able to record four observation sessions using screen sharing of the participants performing actual or example work to create a code set. In particular, two were creating new code sets for their jobs; the other two showing us tools and processes used in their code set development work. One of these presented a code set they were working on but was waiting on feedback from clinicians before doing further work. The other could not discuss current projects for privacy reasons but showed previously published projects.

Many of our participants are atypical in their high level of professional expertise. Many have written papers cited here. Most have published peer-reviewed papers in medical informatics and/or other fields. Several have played key roles as epidemiologists or software developers in the development of tools and distributed research networks for phenotyping and conducting complex observational studies on RWD.

4.4.2 Purposes: Why develop phenotypic code sets?

People work with phenotypic code set across a range of fields and industries doing many different types of work. Our participants work on projects in drug safety, patient safety, clinical quality measurement, informatics (software, standards, or infrastructure), and health economics. Some of our participants work in organizations or departments whose primary purpose is drug safety or patient safety. Others work for universities (and/or teaching hospitals) or firms that provide technical or research services, but their projects involving phenotypic code sets also fall in one or more of these areas.

People also work on phenotypic code set-related tasks within many different types of organizational arrangements: providing regular or ad hoc services for "clients" within or external to their organizations; or as part of their own work. Projects may be funded by grants, contracts, or organizational budgets. They often involve collaboration across organizations. They may be oriented toward publication, clinical or public health decision-making, administrative goals, or some combination of these at once.

Projects differ widely in their access to time, money, expertise, data, and other resources. And methods for constructing and evaluating phenotypic code sets differ widely depending on whether it is meant for a single analysis, repeated analyses with a single database and terminology, or repeated analyses across different databases and/or terminologies.

4.4.2.1 Drug safety

A major part of FDA's work regulating approved drugs centers around analysis of RWD—in fact, the terms real-world-data and real-world evidence were introduced to the field by FDA authors[1]. Drug safety monitoring includes *signal detection*—finding previously unknown adverse events caused by approved drugs being used in the general population—and *signal verification*: determining the relative risk of a specific adverse event associated with a specific drug as compared to alternatives. Drug safety studies are performed by regulators (e.g., FDA in the U.S., EMA in Europe), pharmaceutical companies, academic researchers, health systems, clinical research organizations, and others.

The extended quote here touches on a number of concerns that come up in crafting a code set for a drug safety study:

> …keeping with the anticonvulsant, we might export that data and group it by brand name, generic name, route and maybe some ETC category. We group them up. Then we review them with […] clinicians to see which products you actually want included. Anticonvulsants is almost never going to be the end of that query. What I was showing is actually the beginning of a long process, because what will we do? We'll say, "Well-"
>
> I don't want bulk product. I don't want injectable. I only want the 20-milligram tablets. There's all of those decisions. If you're doing a study in infants, are you going to look for tablets or are you going to use the sublingual? You can choose to use them all or you can choose to be more specific. That's a decision that gets made by the […] team, because it depends on the purpose of the query. To push that example, you're looking at kids under six months old and you want to know what anticonvulsants they're using. If all you're doing is to try to get a general sense, you might just include everything. Does anyone under six use anticonvulsants, out of the million kids that are that age? Is it 10 kids or 100 or 1,000? Then maybe you don't worry about the bulk products and injectable versus tablet versus dosing, anything. You just want to look. If you're doing a comparative study, now you're going to want to be really specific.
>
> There's a version of it as a combo product that's used for weight-loss. Do you want to include that one? A few, but I'm not making that decision for you. I'm going to present what we have and then the researchers make the decision of what's included and excluded. It might even require some querying to see how the product is used before you make the final decision. (*P04*)

4.4.2.2 Patient safety

Patient safety analyses look for instances or rates of medical errors, healthcare-acquired infections, and other iatrogenic threats to patient health. It is performed by many of the same organization types that do drug safety monitoring; though, in the U.S., it is the purview of AHRQ rather than FDA.

4.4.2.3 Clinical Quality Measures

Clinical quality measures are standardized algorithms (each including multiple value sets) for measuring health provider compliance with clinical guidelines. Medicare reimbursement rates can be tied to providers' implementation and execution of these algorithms, reporting to CMS,

and meeting certain quality levels. *P02* is a manager and researcher at an organization that develops CQMs used nationally. Here they are speaking generally about concerns arising in value set development in this environment.

> …sometimes the measure wants just a very generic definition…But other times if you just want postpartum depression or if you don't want postpartum depression in the general value set because it's a separate set for a specific population, you know, it spirals from there. So if the measure wants everyone who has any indication of depression, usually you can include everything, but very few measures want that broad a definition. Most of them want a very specific definition and we'll either exclude our recat or reclassify those subpopulations differently. (*P02*)

The following example speaks to the purpose and use of both the value set being developed and the repository of value sets made available by VSAC. *P16* works on and publishes about both phenotype development and value set/CQM development. Here they are describing construction of a value set for depression that will be used in a patient-reported outcome measure, a kind of CQM based on both RWD and patient survey results. They have downloaded hundreds of value sets and written Python code to itemize their shared and different codes as part of the process of assembling and evaluating their own.

> …multiple value sets; so the eCQM includes a value set for depression screening for patient refused, for bipolar, for depression diagnosis, grouping value, et cetera, so every single one of these is a value set that appears in one of those eCQMs… (*P16*)

After several minutes discussing this project, the interviewer wonders how typical it is for CQM development and asks if other CQM developers go through anything like this kind of process.

> **P16**: Not entirely. I feel that they go through steps but maybe not utilizing VSAC as much as they could. That is very anecdotal on my part though.
>
> **SG**: Because they'd have to write a script and they're not equipped to do that?

> **P16**: Not even writing a script per se. Even though NLM has made it clear that it can be used for any purpose, but this is very much promoted within the quality measure development community. […] We don't see value sets discussed a lot in the phenotyping literature. It is not common for someone to provide a phenotype algorithm that says and here is the VSAC OID for the ICD9, ICD10, and SNOMED codes that you need to use. And again, this is purely … My opinion is I think that there's benefit if we could move more towards a direction where it isn't some type of centralized thought, and it's appropriately annotated so that we can look at reuse or even if people don't agree on it, just that we have this provenance and link.

4.4.3 Two observation sessions

P14, an epidemiologist working in the drug safety division of a large pharmaceutical company, uses the OHDSI collaborative's open-source ATLAS web interface [7,116]. In this environment, they have access to a large set of interlinked vocabularies, and dozens of large, vendor-supplied databases of deidentified patient data from all over the world.

In the following example they need three code sets for three different preparations of a drug made by their firm. The drug is an injectable sold under three brand names (in the U.S., other names elsewhere): one lasts for one month, the second for three months, and the other for six months. Codes in patient data tend to be based on the ingredient name rather than the brand name. Since most of the RxNorm drug codes for all three formulations will be labelled with the ingredient name and not the brand name, assigning drug codes to the right formulation will be a challenge.

They begin to establish and refine a single code set for the ingredient, narrating as they go along.

> What I at least do is start with looking for ingredients, which is always a pain in the butt, finding the ingredient thing in here. I might even have been looking in the wrong folder. Where is it? It's hidden from me. Precise ingredient, that's not

> what I want. There it is. Oh, I just saw it go by. There we go. Okay. So then I would put that into my concept sets, and I would start taking a look at what's underneath of it, like on RxNorm at least.

Another participant, P43[12], an informatics professional and researcher at a university in the UK, is making a code set to represent death. They are using GetSet [27,32], which, unlike the ATLAS/OHDSI example above, searches a single vocabulary, Read Codes, because the code set they create will only be used on UK data coded in this vocabulary.

In P14's process, they frequently consult the patient counts displayed along with the vocabulary terms to assure that they are looking at codes that actually get used in their data—as well as for hints to help them distinguish the three brand names since they have some sense of the relative sizes of the populations using these drugs.

P43's data, unlike P14's, is not deidentified. They do have patient counts for codes, but those are not part of the GetSet interface. Fully deidentifying patient data is difficult and costly. Few of our participants have the luxury of working in the kind of data-rich environment P14 enjoys.

Before they begin constructing a code set to find indications of patient death in the EHR, P43 opens up the tools they will use:

> I guess I've got a few things I would do when making a code set. First of all, there's GetSet. I'll also open up a couple other things, because I'm doing it for READ codes. There's an app that the NHS provides, with NHS Browser...
>
> This is the READ code version two browser. It basically gives you all the codes in the hierarchy. It's not the easiest tool in the world to use, but it's sometimes good, just sanity check things. Okay, so I'll leave that for the time being. The other thing that I've got is … I've got a database, so the main data set that we usually

[12] P43 requested that their identity be included with quotes; it is Richard Williiams, https://orcid.org/0000-0002-0920-1103.

> use are patient records. I've got the, all the codes and the frequency of those codes that are used within the system.

Nevertheless, the gathering of codes begins with a simple string search in both cases.

P43 began by searching for "death" and retrieved 142 coded terms containing that string, as well as another 79 "unmatched descendants"; that is, terms not containing that string but being related (as descendants, synonyms, or siblings) in the Read vocabulary of codes that did.

The remainder of R43s process continued based on inspection of those 221 terms. When they found a code that seemed not to indicate patient death, they would enter a string to match it as an exclusion term. For instance, for "family bereavement" they entered "bereavement," which excluded any term containing that string and all their related codes. They discovered that the exclusion of "bereavement" also excluded "Sudden infant death." They would need to do follow-up research to determine if the infant death term would be used in the infant's patient record as well as in their parent's patient record.

4.4.4 Variability of effort

We found significant variability in code set development practices and norms along numerous dimensions:

- The formality and rigor of code set development workflow and validation
- The types and degree of expertise of those managing code set development and the larger analytic process
- The degree of integration between these processes and their resulting components, such as study protocols, phenotype, cohort, and analytic variable definitions

- The information and resources used in code development, such as terminology resources, code set repositories, CDWs and software tools and platforms
- The effort spent in exploratory analysis of patient data and terminology resources
- The formality and types of reports generated, if any, of assessment processes and measures.

The following quotes encompass the range in the formality of validation we saw in the answers to survey question 16 ("How do you verify that you have selected the best codes for representing a clinical concept in your analyses?"):

> 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)
>
> 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)
>
> We are usually given the value set (P12)
>
> By my background knowledge (P23)

Strategies for gaining confidence in code set accuracy vary widely in their apparent formality and rigor, from systematic clinician chart review to analyst spot checking of patient records, from Delphi techniques to vaguely defined sanity checks based on vocabulary review or expert knowledge.

The variety we see in our participants' answers complicates Williams' more formal ideas of a validation phase, which includes activities such as the measurement of statistics such as sensitivity, specificity, and/or positive predictive value (PPV) and the reporting of the methods used in construction and validation [17].

4.4.5 Reporting and sharing standards

There is a wider literature on RWD reporting guidelines [5,24,103,107,117–119] with specific recommendations on what aspects of study design, cohort selection or phenotyping, and code selection and the validation of these should be reported. This literature is concerned with reporting individual studies in the literature and does not generally address questions of how study elements like phenotype definitions and code sets and their validation could be shared to meet the needs of potential re-users.

Technical requirements for reuse are covered in "The FAIR Guiding Principles for scientific data management and stewardship" [120], FAIR being the acronym for Findable, Accessible, Interoperable, Reusable. P07—a professor of medical informatics specializing in semantic interoperability, terminologies, information modeling, and standardization—offered as a best practice to facilitate discovery and reuse of code sets, "Representing them as FAIR objects." And as a change they would like to see in the design of code set management tools, "Enable use of and contribution to repositories of FAIR objects."

Beyond the requirements for making reuse of code sets technically possible—already accomplished by the VSAC and other tools—meaningful reuse will also depend on potential re-users of a code set being able trust its quality and fitness for their use, which will require access to evidence regarding its validation and purpose. Alper, et al., 2021, "Categorizing metadata to help mobilize computable biomedical knowledge," expands FAIR to FAIR+T to encompass trust [104]. Though that paper offers example metadata for value sets and other computable objects, it does not address the difficult problem of sharing code set validation evidence.

4.4.6 Quality: What's a good phenotypic code set?

Though it seems everyone would agree that phenotypic code set quality matters, and there is a difference between a good code set and a bad one, and that making a good one requires work and expertise, there is no objective, context-free standard by which code set quality can be judged; the best, most rigorously validated code set will be good for some uses and not for others. The need for differing code sets may arise for various reasons, such as differences in analytic goals.

> Code sets are always context specific. There is no such thing as diabetes in a 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)

> We don't have a gold standard. The problem is how to assess the value and performance of a value set for a given purpose…If I'm writing a grant and want to impress a funder, I want a sensitive value set, but if I'm running a study, recruiting patients, I want a specific value set. (P12)

The particulars of containing phenotype algorithms also affect code set requirements.

> I don't think you can separate notion of concept set from cohort, because concept sets are devoid of data, but its application of concept sets, alongside temporal logic, that becomes instance to find persons in dataset. (P05)

And phenotypes and code sets perform differently on different databases and patient populations.

> Every database shows different patterns (P05)

> Not looking across multiple databases is the equivalent of burying our heads in the sand. (P05)

Clinical code set developers often lack access to patient data that their code sets will be used with [17].

Not all participants had access to data, even when they expected to.

> Early on in project we didn't have the data. Making the algorithms blind to the data is useless!! (P31)

> We thought we had done a good job, and we had done a shitty job. We found out because we hired a fellow and asked him to come up with scenarios with patients who would have false neg and pos. There were temporal issues, we needed three encounters for diabetes, but had a patient with just one encounter. He found issues with each of the algorithms concerning to us. And we didn't have the ability to run stuff yet – we didn't have data yet. (P31)

Our participants used a variety of tools, some made for general-purpose programming or analysis (Python, R, SAS), others built to support health records research and providing dedicated code set management features, such as ATLAS or Apelon's commercial DTS [121]).

Some participants working on CQMs outsourced code set specification to consulting companies or took them ready-made from clients or published observational studies.

4.4.7 Two classes of data and analysis: semantic and empirical

In the terms of some of our participants, the focus on meaning is called *semantic* and the focus on use is called *empirical*. When our study participants spoke of empirical data, they were referring to patient-level records; when they spoke of semantic data they were referring to codes, definitions, and relationships within and across controlled medical vocabularies. This distinction came up in P05's survey response:

> Lexical search,[13] semantic exploration (navigate OHDSI vocab), empirical assessment thru characterization, and clinical expert review.

Through the survey data coding process, this distinction evolved into semantic and empirical, the two biggest categories grouping information sources for code set construction and evaluation.

[13] The distinction here between lexical and semantic is simply between string search on vocabularies, and navigation of their hierarchies and other structures. For simplicity, we consider both to be "semantic".

Other categories, such as code sets from publications or repositories eventually ended up subsumed under semantic. Later it became clear that this was a productive categorization for thinking about structural dimensions of study contexts, code set reuse, software design, styles of exploratory and confirmatory evaluation.

For our use, we give the terms semantic and empirical precise, practical definitions, which may seem idiosyncratic but preserves the denotative meaning understood by our participants, clarifies some edge cases, and concretizes the distinction to provide further practical benefits.

Empirical data (for our purposes) are about patients and their relationships to codes and code sets. The reliance of empirical techniques on patient-level data limits their scope, because empirical data are not always available or usable during code set development due to privacy and other concerns.

Though empirical data and techniques may be indispensable for validation depending on analytic and reporting requirements, patient-level data are generally private, protected, institutionally idiosyncratic, missing data elements that may be needed, expensive and cumbersome to access and use. These data can seldom be fully deidentified and generally qualify as protected health information (PHI). There can be crucial differences in the data elements available and how these are structured and accessed across institutional and database contexts. While code sets target coded data, empirical validation may require clinical notes, lab results, images, and other non-coded data. Even if these data will be available when the full study is run, they may not be available to code set developers. Some data constraints, limitations, or complexities may not be discovered until deep into a project. All told, the use of empirical

techniques is often encumbered with high resource demands and legal, institutional, and technical hurdles [110].

Patient and occurrence counts play important roles in code set development and evaluation. Though generating them requires access to patient-level data, they can be separated from PHI and used more easily by code set developers and code set authoring software.

Semantic techniques rely on clinical terminological knowledge and data from controlled vocabularies, groupers, and existing code sets. Though advanced clinical terminology software and resources are not always at hand, code set developers will always have access to at least basic code lookup and navigation of the vocabularies relevant to their projects.

Semantic data (in the form of digitally encoded vocabularies, groupers, code set repositories), though they may be encumbered by licensing constraints and complexities (e.g., around version synchronization), are not burdened with the same kind of privacy issues that are always present with patient data. Though medical terminologies may differ greatly in structure and classificatory principles, dealing with these differences is mitigated for users of curated and harmonized vocabulary systems such as UMLS [122], the OHDSI/OMOP vocabulary systems [86], and CDMH [123,124].

The small portion of participants discussing empirical validation may be evidence of the difficulty of performing it. It is also striking that more technically proficient participants who discussed writing analytic code (in SQL, Python, etc.) all had some degree of access to empirical data, but—in the service of code set development—only used semantic data. Semantic data, in contrast to empirical data, can be distributed and used without exposure of PHI.

We call attention to the association of semantic with sharable data and empirical with protected data believing that it sheds light on challenges involved in designing code set management tools and techniques relying on empirical data, as well as on opportunities to make more effective use of semantic data in tool design.

4.4.8 Semantic techniques and data

4.4.8.1 Expert knowledge

It takes some clinical expertise to frame an analytic question around RWD and to make sense of the answer, and it takes some informatics and vocabulary expertise to implement an analysis and choose more or less reasonable codes. Experts may be engaged in various ways: individually, on panels, by checking their judgments against each other. Though only half of the responses to the validation question (Q16 in Appendix A. Survey questions) explicitly mentioned expert input or review, experts are always involved in one way or another.

> We have a clinical expert review the concept sets. (P26)
>
> Review with informaticists and subject matter experts (P08)
>
> We discuss with our coding panel, a group of experts that give us advice and feedback. (P21)
>
> Oversight with terminologists, clinicians (P19)
>
> Depends on domain. Usually verify with clinical SME when available, or conventions within a network (P44)
>
> Manual review with knowledge of the coding system and domain is critical. I usually find a trusted expert (ex: pharmacist or ontologist) and then review the code sets manually based on the intent. Find a trusted expert who isn't trying to obtain publication for the effort - the goal of publishing inherently biases the work. The person should be a subject matter in the domain (ex: pharmacy) as well as the terminology standard (ex: RxNorm). Engage them and review. (P24)
>
> Examine the literature for validation studies. Take code lists to clinical and coding expert panels for review. (P41)

Although expert consultation is generally considered necessary, it is no guaranty, at least according to P31, of validity:

> After we wrote the first set, we shopped it around to physicians (specialists) to see if there were relationships that we could use to increase the sensitivity and specificity of algorithm.
>
> Chart review would have been better. Sensitivity analysis would have been better than nothing, but we couldn't even do that.
>
> Get two people to elect the codes then have the final decisions independently reviewed …. and it is still error prone.

Expert review can occur along a continuum of formality and rigor. At the more formal end, we had one participant describe systematic code set development using a Delphi-like process with inter-rater review.

4.4.8.2 "Authoritative" sources

Besides consultation with or review by terminology and domain experts, participants discussed using published and other public or available sources: existing code sets found in published literature, code set or phenotype repositories, and published code groupers. We could describe both types as "validation by authority;" their credibility is based on the authority or source's reputation.

> Examine the literature for validation studies. (P41)
>
> Previous published results (P01)
>
> Prior literature, technical reports (P36)
>
> Verifying against the current billing practices (claims) or validated phenotypes (P06)

Expert knowledge and existing code sets can both serve as sources of authority; they serve to bolster confidence in the appropriateness of the codes they suggest or contain.

4.4.8.3 Semantic data

Semantic data play an indispensable role in code set development. Data representing vocabularies relevant to a code set development task can be accessed using a wide variety of formats and tools: CSV or other raw files; physical books; relational or hierarchical databases, triplestores or other ontology resources; clinical terminology services frameworks and APIs such as CTS2 [125,126], FHIR's terminology services, or Monarch's BioLink [127]; or graphical or tabular browsers like https://browser.ihtsdotools.org/, OHDSI/ATLAS vocabulary and concept set tabs [116], or proprietary tools from IMO, 3M, and other vendors.

Software tools are also available particularly to help code set developers discover codes related to the ones they start with: Term Sets [27] and PHenotype Observed Entity Baseline Endorsements (PHOEBE) [128]. PHOEBE functionality has recently been added to OHDSI's ATLAS concept set editor.

Semantic resources fall into three general types: vocabularies, cross-vocabulary mappings, and code sets (referred to as value sets in the clinical terminology services context). In addition to individual vocabularies, some systems include multiple vocabularies and mappings between them, such as UMLS, OHDSI/OMOP's vocabulary tables, or the Mondo Disease Ontology [129]. Cross-terminology mappings are also available from other sources [see 42, p. 447, footnotes 4, 9, and 10] and in other formats [130,131]. Code sets, as discussed above, are available from code set repositories such as VSAC, ClinicalCodes [30], or the OHDSI/ATLAS or N3C concept set editors [10,116,132]); published papers that follow RECORD and other data-based observational study reporting guidelines [5,24]; previous projects available to the code set developer; and groupers.

Semantic data (as opposed expert review or empirical analysis) tend to be seen as resources for constructing code sets rather than validating them. The development process begins there.

> Terminology browser should be first point of reference for direct reference codes as it is the best source. For broader clinical concepts that involve multiple terminologies, clinical quality measure value set directories are a good source of curated codes that represent valid concepts. (P02)

> Needed to start with semantic criteria because I can't get data without a code set. (P34)

Source vocabularies can be probed using string searches and exploratory navigation of their hierarchical structures.

> Finding common ancestor codes to help create consolidated concept set definitions. (P33)

Vocabularies are large and have complex structures. Ontologies and collections of cross-mapped vocabularies can be significantly larger and more complex. They are packed with codified knowledge, embedded in the semantic graph formed by codes and relationships between them regarding causality, physiology, function, or etiology, in addition to simple subsumption relationships. Specialized software is required to "navigate" them effectively.

> We have a software tool that helps search the terminology. (P30)

> Having the ability to compare concept sets in a visual way could be useful. (P33)

Most standard medical vocabularies have their own dedicated terminology browsers. Browsers for terminology collections, such as UMLS and the OHDSI vocabulary system, are also available. Some participants find vocabulary browsers challenging to use at times.

> Have used OHDSI/ATLAS and Athena browsers. Compared to SNOMED—ok. . ., major drawback is ordering of results, generally alphabetic and not closest matched (P07)

In a passage above P14 says, “this is always a pain in the butt, finding the ingredient thing in here.” P17 describing how they assemble code sets for data requests in i2b2 says,

> Gotta drag every code individually. It’s not easy to exclude subsets of codes. You can’t do it in bulk.

Several participants spoke of using existing code sets from literature, repositories, code grouping systems (e.g., CSS, MEPS, DRGs), and other sources.

Semantic or empirical evaluation can reveal problems caused by new vocabulary version releases. P34 here is talking about the quality of existing code sets they might use for their own projects.

> I’m not sure- but one needs to know the purpose (research or clinical care), how they were validated, results of validation, and how they will be maintained. All one needs is for a billing/reimbursement process to change and a new ICD10 or CPT code will be used, that wasn’t used in the past and wasn’t relevant, and then the value set needs updating. In a perfect world, it would be nice to be able to re-validate on demand with test data and one’s own data. (P34)

Several participants discussed their use of vocabulary browsers in the process of constructing code sets. Some used vocabulary browsers provided by the vocabulary maintainer, others used third-party browsers such as the OHDSI ATLAS concept set builder [116] or OHDSI’s Athena [133] tool for browsing and downloading vocabularies, one used the vocabulary tree built into i2b2 [134]. All these tools allow hierarchical navigation of vocabulary structures.

P14, the drug safety epidemiologist discussed above, narrated some of the difficulty of using ATLAS to search for generic or brand names for a particular drug:

> Precise ingredient … that's not what I want … there it is … oh, I just saw it go by…

And later, on a related task:

> I can see the brands, but I actually can't really use ATLAS to split them up very easily. So, to be honest, I would probably need to go out and write some SQL at this point.

P43 discussed using GetSet [32] to iteratively refine a code set using Read Codes: an initial string search immediately populates the code set with terms containing the search string and with all the descendants and synonyms of those codes. If the resulting list contains terms inappropriate to the code set, search strings that match them can be added to an exclude list.

Another participant wrote Python scripts to download and compare a number of depression code sets from VSAC to find the most commonly used codes and to consider whether less commonly used codes should be included in a code set meant to cover depression generally.

> There are several value sets that define a diagnosis of depression, and so my first questions are, okay, why are there multiple value sets that do this? What are the differences? And then, as we get further into the process with the clinical experts on using, reusing, or expanding a value set I want to understand what that overlap is...
>
> [I]n the ideal world we'd resolve to an existing value set that we could reuse but there's many value sets for similar concepts for a reason, and so it may be that we do need to create a new value set based off of an existing one… I'm sitting here wishing that there was an explicit acknowledgement by one of these authors that, oh yes, I know that this bipolar disorder grouping value set existed. I was aware of it. We reviewed it, and we created this one because dot, dot, dot. (P16)

Though many participants mentioned starting code set construction with string search in vocabularies, none considered it sufficient by itself for producing a code set. One participant suggested, however, that many researchers do think it sufficient:

> Sadly, most investigators still use string pattern matches for their work and don't do ANY analysis of the coded values…I usually have to do cursory checks of the resulting codes and quite often find obvious issues…manual review with knowledge of the coding system and domain is critical. I usually find a trusted expert (ex: pharmacist or ontologist) and then review the code sets manually based on the intent... (P24)

4.4.8.4 Empirical data in the service of semantic search

Empirical data can be important in the process of code collection as well as in validation. Multiple participants mentioned wanting to use patient data to identify additional condition codes by examining the morbidities of patients taking drugs indicated by those conditions; or otherwise discovering semantic or clinical relationships by exploring frequently occurring conditions or treatments once they had begun to delineate a relevant patient population. This, like sensitivity analysis appeared to be more a wish than an established practice; though P14 did say they had used drugs to find conditions and vice versa.

4.4.9 Empirical techniques and data

4.4.9.1 Chart review

With the exception of patient and record counts, when participants discussed empirical analysis, it usually included some amount of chart review. We saw little evidence that their chart reviewing practices were highly systematic or geared to building a gold standard. P16 mentioned performing "random chart review to help detect if presence of code indicates disease."

P34 suggests that comparison with other code sets may be good for sensitivity but that chart review can help in achieving specificity. P28 describes their validation process as,

> Multiple Physician review of code lists and their opinion of appropriateness. 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.

P31 would have followed formal empirical methods, but it "would have been cost prohibitive to do chart review even if we had the data."

Although every one of our survey respondents (36) report that their studies use patient data, of the 32 who answered the open-ended validation question, only 9 indicated using patient data for validation. These include those discussed above who use chart review.

There are also cases where empirical data is being consulted, but possibly in a more casual way than chart review:

> I do some analysis of result set produced by queries with a given value set, using descriptive statistics and visualization. Sometimes I will also look at data missed by the value set to see if there are additional revisions that I should make. (P17)
>
> Manual reviews. Check against public dataset (if exist). (P09)

Four participants attested to comparing results across multiple databases.

> We benchmark across datasets in the federated analysis…We need tools that provide] better visualization of relationships in vocabularies that show the prevalence of the codes across the network is useful. […These] would help to define concept sets that work in multi-database studies. (P10)

Our participants discussed a range of empirical techniques for testing their code sets: doing spot check inspection of patient records, more systematic chart reviews, and use of record counts, either for individual codes (with or without descendants) or for whole code sets. The following quotes come from faculty and staff at academic medical research centers with their own CDWs.

> 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)
>
> We benchmark across datasets in the federated analysis. (P10)
>
> Test in real world data. (P41)

4.4.9.2 Reference standards and formal validation practices

While our study participants report a wide range of formal and informal code set validation methods, the most rigorous method for empirical validation is by comparing results against a reference standard, which could be a registry or created by systematic medical record abstraction (MRA), a systematic process of chart review conducted by qualified clinicians or trained abstractors [5,19,24]. Reference standards allow the calculation of sensitivity and specificity or positive predictive value (PPV). *Rethinking Clinical Trials* [19] bluntly states,

> Estimation of validity requires a gold standard, defined as the best classification available for assessing the true or actual phenotype status. Assessment of a gold standard is a resource-intensive process that requires careful manual review of current and historical individual patient data.

The only validation method capable of yielding a full 2x2 table (see Table 3) and producing sensitivity and specificity statistics is by comparing matched records against a reference standard – a sample of records already tagged as presenting the phenotype or not. This approach is offered as a recommended guideline in the RECORD Statement [5, p. 4, items 6 and 7] and in Callahan, et al. 2020 [24, p. 10]. The RECORD statement further recommends expunging codes used in the phenotype or code set from the human-reviewed charts to avoid bias. This is the current normative standard, as far as there is one, for RWD studies and computable phenotyping—though its appropriateness has been heavily critiqued and its and execution may be rarer than expected.

Many of our participants mentioned chart review, sometimes implying such a formal process, other times implying no more than cursory sanity checks on a handful of records; usually by a clinician, though sometimes by an analyst without clinical training, such as an informaticist or epidemiologist. Even when conducted according to recommended guidelines, MRA is a

complex, multifaceted task, prone to subjectivity, with “high and highly variable discrepancy rates.” [135]

4.4.9.3 Refining versus validating code sets

If a reference standard is available, the set of records tagged in it can be run through the phenotype algorithm and a 2x2 table (Table 3) can be constructed. With all four cells of the table at hand, sensitivity, specificity, and PPV can be calculated and reported.

Table 3. 2x2 table with MRA-generated reference standard

	CONDITION PRESENT IN DATABASE	CONDITION ABSENT IN DATABASE
ALGORITHM POSITIVE	True positive rate calculated from comparison to reference standard	False positive rate calculated from comparison to reference standard
ALGORITHM NEGATIVE	True negative rate calculated from comparison to reference standard	False negative rate calculated from comparison to reference standard

Without a reference standard the only digital evidence of the clinical phenomenon or phenotype being present or absent is what can be found in the CDW, either by ad hoc forays into the data or by running the code set through database queries or the algorithm the code set is being designed for. There is a less resource-intensive but systematic way of measuring code set accuracy than creating a full reference standard: performing chart review or MRA on a random sample of the records matched by a code set or phenotype to identify false positives. This procedure allows reporting of PPV, but not sensitivity or specificity [22].

Williams [17] highlights two particular code set validation processes: (1) ”internal validation typically with a clinician…when the final code set is examined to confirm that all the codes included are relevant to the concept of interest. This is an important step to reduce type I errors,

where an incorrect code is wrongly included," and (2) "creating a list of synonyms, using a code hierarchy and searching iteratively…, an important part of the code set construction process, [for] reducing type II errors, where a valid code is incorrectly omitted [17, p. 8]."

Type I errors, the inclusion of inappropriate codes, tend to increase false positives in the results of patient selection, and type II errors, the omission of appropriate codes, tend to increase false negatives. One participant mentioned that when codes are added to a code set to correct type II errors, the intent is to transform false negatives into true positives, but the new codes may also transform true negatives into false positives.

Without a reference standard to identify false positives and false negatives, however, misclassifications must be discovered by various semantic or empirical approaches to identifying them. With sufficiently sized random chart review, researchers may feel confident they have identified false positives. With true and false positives at hand, PPV can be calculated and reported. This approach will not help in identifying false negatives, as illustrated in Table 4.

Table 4. 2x2 table with random chart review of positive results

	CONDITION PRESENT IN DATABASE	CONDITION ABSENT IN DATABASE
ALGORITHM POSITIVE	True positive rate calculated from random chart review of positives	False positive rate calculated from systematic random chart review of positives
ALGORITHM NEGATIVE	False negatives unknown; finding condition amongst huge set of negatives probably not feasible.	True negatives = all negatives minus (unknown) false negatives

If neither reference standard nor systematic random chart review are used, none of the 2x2 table cells will be known (Table 5). Even supposing there is enough evidence somewhere in the database to support a determination of whether any patient has or does not have the condition, the evidence available to the code set developer consists of (1) what the algorithm (or code set)

matches, and (2) whatever else they can determine or surmise by the variety of techniques they are able to deploy.

Table 5. 2x2 table without systematic chart review

	CONDITION PRESENT IN DATABASE	CONDITION ABSENT IN DATABASE
ALGORITHM POSITIVE	True positive unknown	False positives unknown; may be discovered by unsystematic chart review (spot checking)
ALGORITHM NEGATIVE	False negatives unknown; unlikely to be discovered	True negatives unknown

Editing a code set consists of adding and removing codes. Codes are added in order to correct type II errors, eliminate false negatives, and increase sensitivity, but may inadvertently introduce false positives. Codes are removed in order to correct type I errors, eliminate false positives, and increase specificity, but may inadvertently introduce false negatives.

If false positives or false negatives are discovered, the developers will presumably modify the code set or phenotype. If no systematic empirical validation is performed, its authors may assume, with or without justification, that correcting these errors will improve its accuracy, but they will not be able to quantify that improvement.

Nevertheless, iterative improvements in code set quality are common in real-world practice. Some experts have more confidence in rigorous refinement efforts than in the statistically reportable results of MRA-generated reference standards.

Unfortunately, these efforts are seldom documented. When we asked participants about the value of documenting refinement efforts, they expressed enthusiasm, but affordances for accomplishing this can be found in few if any available tools.

4.4.9.4 Sensitivity analysis

Several participants mentioned or discussed sensitivity analysis, though this appears to often be aspirational rather than a systematic or routine practice. Some particularly expressed a wish for software that could facilitate sensitivity analysis.

> Tradeoffs between using one value set or code versus another should be visually displayed. Understanding the implications or whether or not a value set is used in CQMs, and if so, exactly which ones, and how it might impact the CQL or expression-logic. Metrics should be explored for synonymy, and concept orientation to see if value sets are "complete" or not from the taxonomy perspective. (P28)

> Assess the impact of including/excluding the codes on patients included in clinical quality measures, heuristically evaluating whether siblings/synonyms etc. should be included or are appropriately excluded. (P29)

> By tools, what would have been most helpful, to vary the code sets and data elements and vary the logic, and then make Venn diagrams of patient sets. And once i had that ugly flower of Venn diagram, I'd like to look at the outer regions and see why they fell out. (P31)

Some participants who had access to patient counts, did basically perform casual, iterative sensitivity analysis by making changes to code sets and then applying them to data or running the analysis and checking the results. P12 described a practice of sensitivity analysis in which the entire study was run using a full range of plausible code sets. If results remained stable across these, any differences in cohort composition would be considered immaterial.

If prevalence of the phenotyped condition is known for the population represented in the counts, a phenotype that matches the expected number of patients *might* be accurate—but it could also be a coincidence of having equal numbers of false positives and false negatives. On the other hand, an unexpected result is a strong indication of a problem.

4.4.10 Patient counts for codes or cohorts

In addition to validation methods based on patient-level data, several participants discussed term usage counts as important to their code set development processes.

> Compare them to published literature [to] determine the presence/absence of codes in our target datasets (P33)

In answering Q16, P05 expresses the

> need to show concept counts from across whole international network, empirical comparisons of consequences of changing concept sets on person counts.

Through ATLAS, P14 could view patient profiles and run characterization reports for millions of patients. The only use of the empirical data they mentioned, however, was to eyeball the patient record counts of concepts they considered including in their concept sets.

> **P14**: ATLAS helps a lot because…it shows if you use this concept set, you are going to find a decent amount of people that get associated to this diagnosis.
>
> I usually sort [codes] by how much evidence they have, and I can see that there's a lot of evidence associated to this kind of top-line SNOMED code…
>
> **SG**: Most of the people I've interviewed don't have something like ATLAS. So, they might have vocabulary resources, they might have a database. But having the patient counts and the ability to look at patient data along with looking at the vocabulary is relatively exceptional.
>
> **P14**: Yeah, and actually, the other example that I'd like to talk you through, like the ADHD, that… It's basically critical because what happens in Japan… In the US, we use ICD10CM, right? And actually, it's pretty detailed. But in other countries, they don't use ICD10CM, they just use ICD10.
>
> If you pick a really good-looking SNOMED code but it doesn't actually connect to an ICD10 code, that means you're not going to find anything in your data. So, you always want codes that at least go somewhere that your data speaks, if that makes sense. So, a really great-sounding SNOMED with no ICD9 or 10 codes is worthless.

P14 is making a point about ICD codes here, but that is because their data is coded in ICD, not SNOMED (though SNOMED is preferred in OHDSI tools for specifying code sets), but they

would likely say that even if a SNOMED code linked to an ICD code but that ICD code did not appear in the data, it would be equally worthless.

> And what Atlas does is it actually helps you see that because I could go into the configuration, and here are all my data sets. We created something [that] actually rolls all the counts together, which is kind of cool, so you don't have to worry about them, which one you're looking at. But I also have the Japanese data set. So, for example, I could go in here and see that the US code for ADHD isn't used very often. But its parent is. (P14)

Another participant said their institution has TriNetx, which "only gives you counts."

In order for P17 to do studies at their institution, they explain,

> We need to define value sets as part of our data request process in making a request to the data warehouse. The data we can query through i2b2[134] is very restricted. We only get patient counts and high-level demographics.

An i2b2 query uses Boolean and temporal logic with the records matched by the code lists. At each step of the query, counts of the patients matched are reported. The interface for selecting and adding codes is awkward and slow to use: each term must be found through an indented tree with expanding/collapsing nodes representing hierarchies from all of the available terminologies. On locating a desired term on the indented tree in i2b2's left sidebar, the user copies its code from there to their query.

Given the shortcomings of RWD, few assumptions can be made in their analysis, and it is always wise to check one's code sets and algorithms against empirical data. Unlike patient-level empirical data, counts can be shared publicly.

4.4.10.1 Sharable derived data: an "empirical semantics"

In addition to semantic and empirical, we can define a third category of shareable data derived from PHI but only including term usage counts and other aggregate data as well as evidence and

metadata culled from previous analyses. We could label this category and these data "empirical semantics" with the idea that they will form a supplement, possibly almost an alternative, to vocabularies and authority-based semantic data; a semantics not of intended meaning but of found meaning. PHI cannot be shared, but useful derived information can be. For instance, a study team's code set and phenotype developers could be aided considerably by learning that diagnostic code A in database B only indicated an actual diagnosis of A in 60% of B's patients, while indicating suspected A in 37% of patients, and erroneous data in 3%.

Although term usage counts and other summary data must be derived from patient-level data, the barriers to using summary data can be much lower and designers of code set management tools should consider leveraging summary data even when using patient-level data would not be feasible:

5. Unlike patient-level data, counts do not expose PHI.[14]
6. Since counts can be precalculated, they can be made available separately from patient-level data.
7. A code of interest may occur thousands or millions of times in a database; its occurrence and distinct patient counts are useful and usable pieces of information, whereas review of the occurrence records or chart data for patients with that code tend to be challenging tasks, requiring complex analysis or arbitrary choice of which records to review.

Code set developers, in our observation, are well aware that a record or patient count cannot be taken as an accurate indication of occurrence of the code's clinical phenomenon in the database population. Nevertheless, a large disparity between a term-usage count and the analyst's sense of the prevalence of the clinical condition or event represented by the code in the database

[14] Very small counts, however, can increase the possibility of data re-identification..

population signals the need for further investigation. If counts are lower than expected, this disparity may indicate:

8. That the code in question is not used as expected in the database and the phenomenon is being captured using other codes, or
9. That the phenomenon is not being captured in coded data in the database and either is not discernable in those data or require narrative or other data to detect.

If counts are higher than expected:

10. Many records for this code may have been generated for a single occurrence in the patient trajectory; in this case, distinct patient counts (total or per day) will be more useful than total record counts;
11. The code may be used in patient records as a "rule out" and indicate that the patient was evaluated for a condition rather than diagnosed with it; or
12. The coded concept may be broader than the target phenomenon. (Use of overbroad codes is a particular hazard when translating across terminologies [136].)

Counts for whole code sets are important in the code set collection process: low counts can suggest that codes indicating the condition of interest have not yet been included, spurring further exploration of the vocabulary for relevant terms. Total record counts can be summed across precalculated counts for individual codes, but these totals can be misleading. Unlike record counts, precalculated patient counts *cannot* be summed across codes without resulting in double counting: records for any code in the set must be selected before counting distinct patients.

Although access to patient-level data is needed in order to count the patients matched by a code set being developed, patient counts for pre-existing code sets can be generated and made available, which can be helpful in the code set development process.

4.4.11 Other approaches to code set validation

Some participants believe the best possible approach to reliability code sets and study algorithms is multiple database comparison across a distributed research network, generally

using a common data model and some set of methods for achieving semantic interoperability. Whether or not a reference standard is used to validate performance at each site, by this perspective, study results can only be considered valid if they are consistent across databases. Because databases reflect different populations and different idiosyncrasies in data collection, similar results can be counted as external validation, while dissimilar results can be explored for explanation. Digging into the causes of such discrepancies can uncover institutional factors or population characteristics that shed light on how the results of the study—at each site and overall—should be interpreted. [137]

PheValuator [41], mentioned by two participants, algorithmically produces a "probabilistic gold standard" to be used in validating phenotype algorithms and code sets. It is given two starting code sets, one perfectly sensitive, the other perfectly specific (as well as these can be determined). Patients matched by the specific code set are considered positive cases, patients *not* matching the sensitive code set are considered negative cases. The PheValuator algorithm is trained on patient samples matching each of these and assigns probabilities to remaining patients rather than tagging them as positive or negative.

4.4.12 Exploratory versus confirmatory analysis

When P12 says "We don't have a gold standard…If I'm writing a grant and want to impress a funder, I want a sensitive value set, but if I'm running a study, recruiting patients, I want a specific value set," they mean in part that using a more sensitive or specific code set will have to suffice because the cost of annotating enough records to formally validate phenotypes for every new study would be out of the question.

P04 says you need the one diabetes definition out of a hypothetical 50 that "matches your question, data, and methods."

What seems very clear is that a phenotypic code set's quality or fitness-for-use is dependent on specifics and nuances of the analytic context, calling into question the benefits of reusing other people's validated code sets.

False negatives and sensitivity are a particular concern with gold standards or reference datasets. Positive matches can be examined and verified, but depending on the phenotype's prevalence in the database, the quantity of negatives can be too great to tag a sample large enough to calculate sensitivity reliably. This can be addressed with automated methods like PheValuator [41], but some may be skeptical of automated approaches.

Williams gave particular attention to iterative vocabulary navigation to gather codes and eliminate false negatives. This is essentially an exploratory, not a confirmatory, process. Especially for rarer phenotypes or codes, it can be like trying to find a needle in a haystack. As vital as exploratory analysis is, it perennially gets short shrift in many scientific, academic contexts. Now is not the occasion to rehash the ancient arguments, beyond quoting [138] p. 3,

> Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone—as the first step.

But journals tend to prefer formal methods with quantified evaluation measures, which exploratory analysis cannot provide. It is possible that semantic exploration could discover codes that would not have been included in a gold standard because they did not match the researcher's original code set definition, but when they appear through some semantic connection to the code set being constructed (relationships captured in vocabularies, ontologies, mapping between them,

or similar pre-existing value sets), the researcher realizes that they should have been included in the definition. In this case we could argue that this exploration-enhanced code set would be a "better" fit for the research question being addressed than one evaluated with a gold standard. But it might appear more ad hoc in the final paper as no count of false negatives would be available to allow appropriate reporting of sensitivity and specificity.

Despite the attention of much recent work on empirical evaluation, we believe the most important focus for innovative design work for code set management in the near term will be on tools for semantic exploration and evaluation: there is a wealth of semantic data readily available that has barely been tapped. Browsers exist for ontologies, vocabularies, and vocabulary systems, but not with rapidly interactive interfaces that also use semantic information from available code sets.

4.5 How do people develop phenotypic code sets? How should they?

Williams describes four phases of code set "management": construction, validation, sharing, and reuse [17]. These make sense intuitively: the author constructs the code set, validates it, shares it, and then others come to reuse it. We see from our evidence that there is perhaps no part of the construction phase that would not also be considered part of the validation phase, the process of phenotypic code set development is iterative and not all activities can be easily classed as part of a construction or validation phase. The types of activities performed, and their chronological sequence is highly variable, but we can say that the process of making or finding and customizing a code set generally follows these steps:

- The need for a code set is recognized and defined;
- Codes are added, and they usually come from:
    - Vocabulary string search

    - Vocabulary navigation (by hierarchical or other structures)
    - Existing code sets
    - Expert knowledge
    - Data exploration
- Codes are removed or more codes are added until the code set is judged to be good enough.

If a code set is being developed from scratch, the process generally begins by exploring the relevant vocabularies, but the process can also begin by searching for existing code sets from published studies, public repositories, or a code set might be provided by an internal or external client. In our non-representative sample, around 50% of survey respondents say they use public repositories to look for value sets, and around 75% use value set created by others.

As described above, P16 downloaded and compared all the depression value sets from VSAC. Here they are describing an initial selection they had made from the downloaded codes.

> P16: This really is kind of exploratory where I'm not sure what we're looking for just yet, so I was just playing around with a few different what if scenarios. I have all of the codes, so now I was curious what are all the ICD9 codes.
>
> SG: I'm just curious why you started there rather than ICD10?
>
> P16: Because I'm old and I started with ICD9, so yeah, that's typically my go to just because I had a little bit more exposure to ICD9…This needs to be done with ICD10 and SNOMED, though, for sure. And so, that's also where this is intended more as kind of an interactive portion as opposed to a standalone generated report.

Like some other participants, they describe a back-and-forth process with clients or subject matter experts.

> Now I'm actually waiting on the clinical experts. They are having the discussions of what our entities actually are going to be. Are we going to look for a diagnosis of depression? Are we looking for … How do they define depression is probably the better way to say it, and then my hope is that we can have a collaborative discussion of, "Okay, here is how other groups have defined depression as far as diagnostic codes go. What is your perspective on this?" And being able to show them these are the most popular ones: "Yeah, yeah, yeah, those are pretty obvious. Here are the fringe ones that only a few have identified. Okay, I agree with that. I disagree with that." (P16)

Many phenotypic code set developers have a consistent process and set of tools for constructing code sets. P34, an internist as well as an experienced researcher and informaticist, begins code set construction in a variety of ways depending on their needs for any given analysis, including:

> ATLAS; EPIC groupers; Slicer Dicer; Go to VSAC for say opioids; Start from existing classification from somewhere (P34)

They spoke about a particular study of acne and antibiotics. They used atlas to find:

> antibiotics usually used for acne, excluded things never used for acne. Excluded IV drugs. Looked at oral ingredients, thought about what would be used.

P14 spoke of using ATLAS to find condition (diagnosis) codes by starting with drugs:

> …start not by looking at the condition, but to actually start by looking at the drug that's indicated for the condition…Well, I know this drug is indicated for ADHD. Let me see what conditions get diagnosed the 30 days before the first exposure.

4.6 The Collect, Evaluate, Evaluate, Release (CEER) model

Code set development consists primarily of an iterative construction/validation cycle of code collection, evaluation of individual codes or groups of codes (such as a code and its descendants), and evaluation of the code set as a whole until it is considered fit for use. I developed a process model to give structure and logic to the diversity of practices we encountered in the field. It evolved over two years and many iterations of analysis and writing about it. The final version has four logically sequential phases, Collect codes, Evaluate codes, Evaluate code set, and Release (CEER) and is summarized in Table 6 and Table 7.

Table 6. CEER process model

ONGOING TASKS THROUGHOUT	SEQUENTIAL STEPS	RESOURCE TYPES		
0. Understand - define, - align, - document	1. **C**ollect codes 2. **E**valuate **c**odes 3. **E**valuate **s**et 4. **R**elease	Semantic	Summary clinical	Patient-level detail

Table 7. Code set development practices by CEER stage and resource data type

	SEMANTIC	SUMMARY CLINICAL	PATIENT-LEVEL
COLLECT CODES	Vocabulary search, vocabulary navigation, import from existing code sets, expert recommendation.	Inspect patient and record counts. 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.)	Examine frequently occurring codes in patients with phenomenon for possible inclusion.
EVALUATE CODES	Verify correspondence between code meaning and intended phenomenon: expert advice, vocabulary exploration, research code meaning.		Review patients matched by code(s) to confirm phenomenon.
EVALUATE SET	Review by terminology and clinical experts. Manual or automated comparison with existing code sets. Optimize code set to more parsimonious expression by replacing codes with intensional rules where possible.	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.)	Test code set on patient data, presumably in context of phenotype algorithm. Identify false positives through chart review of matching patients. Identify false negatives if a reference standard is available. Sensitivity analysis of changes in results caused by modification of the code set.
RELEASE	Approval for use by subject matter experts or other decision makers. Describe construction and validation activities and qualifications of approvers.	If prevalence of the physical phenotype (i.e., 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.)	Sensitivity/specificity or positive predictive value reaches predetermined threshold, or informally decide that marginal quality gains do not justify further review. *Generate table 1 for phenotype.* Report description and justification for validation methods used and any resulting statistics. If reference standard was used, report how it was made.

The middle column of Table 6 shows the steps involved in crafting the operational contents of the code set: the codes (or intensional rules to produce them) that will be used to match patient records. In logical order: 1) codes are collected and 2) evaluated individually. Then, 3) the whole set is evaluated, possibly requiring the collection and evaluation of more codes or the elimination of codes included so far. Finally, 4) iteration of steps 1, 2, and 3 ceases and the code set is released to serve its intended purpose.[15]

The first column (Step 0) represents all the attendant work that shapes and documents the sequential steps. The third column emphasizes that there is a range of activities that can be performed at each step and the selection of which ones will be performed in any instance depends on availability of the three resource types.

CEER is a generalization that can apply across the diverse ways that code sets are made, from the least to most formal. It accommodates all the steps that might be performed systematically in the most rigorous process. It accommodates processes that focus on any combination of semantic, empirical, and aggregate data. But it works as a logical rather than strictly temporal or sequential model and can help in understanding even the most cursory or unstructured of processes.

We are not claiming that our participants demonstrated or told us about the kind of structured process steps CEER describes. A code set developer is not likely to think, “First, I will collect codes, then I will evaluate them for inclusion individually or in groups, then I will evaluate the

[15] As described below, maintenance of the code set may be required if it may be used in the future and need to stay in sync as new versions of the vocabularies it draws from are released. Code sets can also evolve by being reused with modifications being shared.

whole code set, after which I shall iterate on those three steps until satisfied." The steps of collecting, evaluating, and deciding may be phenomenologically indistinguishable; that is, the analyst may see a code come up in a string search, recognize almost unconsciously that it is irrelevant, and scroll past it.

But if a code set is created and used or found and reused, the full process has necessarily been logically accomplished: the need and requirements were 0) understood to the satisfaction of the developer, 1) codes were collected, 2) some or all were judged worthy to keep, 3) the set was considered good enough, and 4) it was deemed ready for use. To give an example of an extremely cursory process we have encountered:

1. Requirements were understood implicitly by the code set developer.
2. Codes were identified through string search.
3. The first code in the search results was recognized as appropriate and the remaining string matches were rejected.
4. The first code in the search results was recognized as appropriate and the remaining string matches were rejected.
5. The set as a whole, consisting of a single code, was not deemed sufficient. So, iterating from there:
    a. The descendants of the single code were added.
    b. Newly added codes were all accepted by virtue of being descendants of the first.
    c. The code set was now deemed good enough for its purpose and released for use.

As process modelers, one might ask, why should we belabor the point and classify the string search as collection, the barely conscious brush-off as evaluation, and the scrolling past as rejection? If the model were a mere academic conceit, it would indeed be difficult to justify. The justification becomes clear when the model becomes prescriptive rather than descriptive and the steps are broken out in order to perform them in a way that supports the collection of empirical semantics, of metadata that enables reuse.

Easy as it is to create a code set without thought and effort given to code collection, code evaluation, and code set evaluation, such a set is unlikely to perform well in matching appropriate patient records, and this poor performance may significantly affect phenotype and study results. Worse, these effects will go unnoticed without dedicated attention such as empirical validation, sensitivity analysis, or clinical review.

Beyond describing at an abstract level how code set development is currently performed, CEER is also meant to serve as a roadmap for the design of best practices, forms, and software to guide more rigorous code set development in the future. An early version shown in Table 8 included both observed practices (in black type) and recommended practices (in blue type). A red label appears after recommendations as appropriate: SHARING indicates recommended practice that results in metadata meant to be used by future code set developers; REUSE indicates practices that leverage such metadata in the course of developing a code set.[16]

4.6.1 Understand (Step 0)

While CEER's four sequential steps encompass all the work of constructing and evaluating the functional contents of the code set, those steps are surrounded by other work: understanding what needs to be done, what the code set is for, how it should be defined and described; adjusting that understanding as the algorithmic or other contexts evolve or are better understood; and documenting these understandings and the development process itself so as to facilitate review and reuse of the code set as it is being developed and in the future.

[16] The only version that has been published (as a preprint [20]) is not included here; it splits Table 8 into three parts: observed practices, recommended practices, and an empty template for documenting the process as it occurs.

The work of understanding an analytic code set's requirements—its clinical meaning, purpose, and targeted deployment contexts—begins when it is decided or realized that a specific clinical phenomenon being analyzed will be identified in whole or part by querying RWD for the

Table 8. Process model with recommendations

CCSE Phase	Structured model phase	Non-PHI Can be shared and distributed in software and repositories		PHI Not sharable or reusable; often unavailable
		Semantic Data from vocabularies and code set repositories	**Derived**: Counts, ***metadata from prior uses***	**Empirical** Patient data, registries, chart review; reference standards
Iterative cycle: CONSTRUCTION & VALIDATION	**Code collection** Identify codes for potential inclusion in the code set. *Collect these into a separate holding area to be selected for inclusion or rejection. Do not reject plausible codes prior to collection and evaluation.* SHARING	Vocabulary search, vocabulary navigation, import from existing code sets,† expert recommendation. *Interactive tabular or graphical visualization to explore semantic graph of vocabulary, existing code sets and other semantic data. Capture (auto-generated) summary of exploration paths, search strings, search engine queries, etc.* SHARING	Inspect patient and record counts to support decision making. (Observed technique with current tools.) *Recommended: consider these counts during code evaluation, not collection.* SHARING. *Include metadata from existing code sets† in exploratory interfaces.* REUSE	*Identify codes by co-occurrence; e.g., collect frequently occurring diagnostic codes in patients treated with a disease-specific drug.*
	Code evaluation Select single code or closely related groups, e.g., descendants. Evaluate appropriateness of focal code(s). *Capture reasons for inclusion or rejection of focal code(s).* SHARING	Verify correspondence between code meaning and intended phenomenon: expert advice, vocabulary exploration, research code meaning. *Same or richer UI as for code collection. Potentially relevant codes discovered here should be added to code collection.* SHARING	Inspect patient and record counts. 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. *Enhance exploration UI with metadata shared in the development and reuse of existing code sets,†* REUSE	Review patients matched by focal code(s) to confirm phenomenon. *Capture counts of patients inspected and description of patient features influencing decision.* SHARING
	Code set evaluation Testing and review. *Capture reasons for continuing iteration.* SHARING	Optimize code set to more parsimonious expression by replacing codes with intensional rules where possible. Review of whole code sets by terminology and clinical experts. Manual comparison with existing code sets.† *Automated comparison with existing code sets.†*	*Enhance code set comparison or validation techniques with metadata shared in the development of existing code sets.†* REUSE	Test code set (in context of phenotype algorithm) to identify false positives. Identify false negatives if a reference standard is available. Sensitivity analysis of results changes with different versions of code set. *Capture metadata of process and results.* SHARING
VALIDATION	**Code set acceptance** Formally or informally, when iterative cycle ends, code set is accepted and used in its intended analytic algorithm. *Capture reasons for accepting code set as is.* SHARING	Approval for use by subject matter experts or other decision makers.	*Reviewed evidence from existing code sets† may provide sufficient trust to trigger acceptance.* REUSE	Sensitivity/specificity or positive predictive value reaches predetermined threshold, or informally decide that marginal quality gains do not justify further review. *Capture metadata of process and results.* SHARING
	Reporting (optional) to stakeholders, in published study, etc. *Capture computable metadata throughout code set development.* SHARING	Describe construction and validation activities and qualifications of approvers.	*Evidence from existing code sets† may be persuasive to reporting audiences.* REUSE. *If single, whole code set has been reused, with or without refinement, metadata shared now adds to its body of evidence.* SHARING	Report description and justification for validation methods used and any resulting statistics. If reference standard was used, report how it was made. *Capture report as computable metadata.* SHARING
	Maintenance	Update as source vocabularies evolve; older data may still be coded using deprecated or updated codes, so maintenance can be a complex operation.		

Observed and *recommended* code set development techniques by phase and by information resource requirements

Observed (black, regular type): techniques and activities observed and abstracted into this generalized process model

Recommended (blue, italic type): recommended techniques, not observed, may not be supported by current software

CCSE phase (red, all caps): from 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. DOI:10.1016/j.jbi.2017.04.010. CCSE breaks out construction, validation, sharing, and reuse as separate phases. We combine construction and validation in a general iterative cycle, except for validation at final acceptance and reporting.

† Existing code sets from repositories, publications, groupers; identified by overlapping codes or similarity in name or purpose

presence (or sometimes absence) of one or more clinical concept codes. Understanding does not occur at a discrete temporal or logical point in the code set development process, but throughout. The components of this Step 0, the practices that can or should occur throughout code set development are as follows.

4.6.1.1 Define

An analytic code set will generally have a name indicating its target clinical phenomenon and, as a digital object, the code set may consist of nothing but its list of codes (or intensional rules) and this name. Some systems provide affordances for description and other metadata beyond the code set name (e.g., FHIR, VSAC, N3C), others do not (e.g., ATLAS).

Differences in code sets designed for, ostensibly, the "same" clinical phenomenon may arise from a variety of legitimate considerations as well as from human error or differences in clinical or terminological judgment. Ideally, such differences should be accounted for in the code set's definition.

4.6.1.2 Align

Data analysis and ongoing work on the containing phenotype or study frequently necessitates alignment with and changes to code set definition, and work on the code set can likewise necessitate modifications to phenotype and study. The developer or study team's understanding of the overall analytic task, of the required computational phenotypes, and of the code sets required for these phenotypes, including the code set in question, will be evolving, becoming clearer, sometimes changing significantly. This evolving understanding of the code set's purpose

and context affects its implicit or explicit definition and should ideally be captured in code set, phenotype, and/or study documentation.[17]

4.6.1.3 Document

In our study and in the literature on code set development and phenotyping, the importance of documentation is widely acknowledged, as is the paucity of documentation in practice. CEER catalogs the activities and steps involved in code set development, thereby showing the many specific opportunities for documentation and the understanding, defining, and aligning efforts that support it. In the discussion section we are then able to recommend strategies for increasing documentation in practice.

We did not observe detailed documentation of code sets or the decisions made in composing them in our study, but every step in the code set development process provides opportunities for documentation or metadata capture.

4.6.2 Collect codes for possible inclusion (Step 1)

As the code set's meaning and purpose are understood, the analyst or study team makes a collection of relevant concepts from controlled vocabularies used in the target database(s). Terms can be discovered in target vocabularies through string search and/or navigation of the vocabulary's hierarchical and other structural elements. Or they may be known and contributed by subject matter experts in the target clinical phenomenon.

[17] Such documentation or metadata capture can be particularly valuable at the code set level for reuse purposes since reuse (and infrastructure to support it) is more common for code sets than for phenotype and study algorithms.

The code set composition process often begins by performing a text search in target vocabularies. Code collection could end there, though many code set developers would not consider this sufficient:

> Sadly, most investigators still use string pattern matches for their work and don't do ANY analysis of the coded values…I usually have to do cursory checks of the resulting codes and quite often find obvious issues…manual review with knowledge of the coding system and domain is critical. I usually find a trusted expert (ex: pharmacist or ontologist) and then review the code sets manually based on the intent. (P24)

In practice developers may search for and find codes and decide on the spot whether to include them or not. If they decide not to, in our experience, no evidence remains of that decision having been made. Nonetheless, the identification of codes is a distinct and necessary activity which must logically (if not temporally) occur before keep/reject decisions can be made (in the following step).

We recommend the reasoning behind keep/reject decisions be recorded and made accessible for future code set review, study write-ups, or reuse scenarios. In interviews and follow-up discussions when we mentioned this recommendation, participants endorsed it and agreed that this would be valuable to have if software tools could support its capture without significantly slowing the development process.

Codes whose meaning differs from the target clinical phenomenon may be identified for potential inclusion if their presence generally signals the phenomenon.

A few of our participants mentioned the idea of exploring the clinical data of patients known or suspected of exhibiting the target clinical phenomenon. Then they could identify additional candidate codes that occur frequently amongst those patients, and that might indicate the target

phenomenon in other patients. This approach appears to be more of an aspiration than a current practice, though it has recently become more feasible with the PHOEBE 2.0 recommender system [34].

4.6.3 Evaluate codes (Step 2)

Each of the identified codes or intensional rules collected in the previous step may be assessed in a standalone manner. This involves gathering and reviewing evidence regarding the appropriateness of the candidate code to the code set's requirements before making a *keep/reject* decision.

When multiple people work as a team on code set development, much of the effort involves discussion as to whether a given code will appear (mostly or only) in the records of patients with the phenotype. Development team members might argue, for instance, that a given code can be used as a rule-out, to justify lab orders to test for the condition rather than indicating the presence of that condition.

The discussion may include references to terminology structures, definitions, and the appearance of codes in other code sets. Marshaling this kind of evidence depends on being able to find and deploy it, requiring study team members have access to and competence with vocabulary browsing software.

We were not able to observe code set development and validation discussions amongst team members in our study itself, but we (SG, HL, LS) have been party to many of these in our own work. The arguments of experts in the target clinical phenomenon tend to carry the most weight, though they not infrequently disagree. Even when clinician specialists are present, however,

pivotal contributions may also be made by analysts, informaticists, terminologists, and epidemiologists.

In addition to addressing semantic coverage and accuracy of candidate codes, these conversations serve as an exercise in thinking about the different conditions under which codes can appear and tend to range freely across concerns related to phenotype and study algorithms as well.

4.6.4 Evaluate code set as a whole (Step 3)

There may not be a clear separation between evaluation of the individual codes or rules and evaluation of the set as a whole, especially in the kind of study team discussion just described, but we distinguish these as separate steps for a few reasons.

When a code set is being developed for a specific study and phenotype and clinical data is available during code set development, the code set as a whole can be tested in its intended context. Alternatively, code sets in development can be compared to existing code sets that seem to have the same intended meaning in order to identify overlooked codes, to focus additional attention on the appropriateness of codes left out of other sets, or to consider nuances that come to light by reviewing differences in existing code sets.

Whole code set evaluation provides opportunity to:

- compare against a reference standard if available,
- conduct chart review of matched patients for false positives,
- search unmatched patients for evidence of the target clinical condition to identify potential false negatives,
- if prevalence in the database of the target condition is known, compare that with the positive patient count for whole code set/phenotype,
- conduct sensitivity analysis of how different versions of code set affect results, or

- optimize the code set to a more parsimonious expression by replacing codes with intentional rules where possible, which may provide more clarity or insight than considering it as a long enumeration of individual concepts.

4.6.5 Release (Step 4)

At this step, further work in developing the code set ceases and the code set is released to serve its intended purpose. As we have said, code set development can be completed in moments or in months or anywhere in between. There is no accepted guideline defining a stopping point. The purpose of evaluating or validating code set composition and phenotype algorithms is to improve their accuracy. Insofar as statistical measures can be calculated, they serve to document validation work done. If a threshold for such a measure is set, it provides a criterion for ceasing validation work.[18]

If the code set is being developed and tested as part of a phenotype algorithm in the context of a specific study, developer attention may shift repeatedly between tweaking and reexamination at all three levels—study, phenotype, and code set. The surest way to achieve high accuracy is simply to expend more effort and perform as many methods of analysis and validation as possible.

Release is more likely to be performed as an explicit step when code sets are developed by one set of people for use by others or are otherwise being reviewed or documented as digital artifacts distinct from containing algorithms. It occurs when the study team is satisfied that further incremental improvement will not justify the effort it would require. In our experience,

[18] This kind of measure or threshold should not be conflated with statistical thresholds—such as a 0.05 p-value (itself highly suspect as a guideline but accepted in many settings)—which serve as tests of a hypothesis's validity. A p-value ought only be calculated after any work that could change it is complete, whereas measures of a code set's accuracy are calculated after every modification.

though, release often occurs implicitly, at the unremarked moment the study or phenotype issues being worked on do not require any additional code set tweaks. Even though release may occur unnoticed in practice, however, CEER is a logical process model and release is the logical conclusion of it.

4.6.6 Maintenance

Release is the end of the code set *development* process but is not necessarily the end of its life. If the code set will be used beyond its immediate purpose or context—in later studies, later iterations of the same study, in code set repositories—it will need to be maintained. Controlled terminologies change over time and code sets should be updated to reflect these changes. VSAC and N3C, for example, make this an automated process; OHDSI currently does not. Though maintenance is an important part of a code set's life, a detailed discussion of its complexities is beyond the scope of this paper.

4.7 Conclusions

The models and findings presented here make positive contributions to the theory and practice of value set management and the design of authoring and repository software to support work with value sets.

4.7.1 Theoretical

By bringing together understandings from the value set literature [36,44,29,35,38,39,105,139–142,83,143,144,106] with terminological perspectives, the phenotyping and RWD literature [17,18,21,24,30,90,145–147], and our field study of real-world code set development

practices, we provide a broader, more comprehensive picture of how code sets are made, used, evaluated, and reused than has been previously available.

In our reading of the literature, authors often seem to assume their questions and findings apply more generally than they do. Particularly, they tend to focus on either permissible or analytic value sets, not distinguishing between these two types and implying that they are speaking to value sets generally.

One hopes that future contributors to this field will situate their work more clearly, that they will:

- Specify whether their findings apply to permissible, analytic, or both;
- Indicate whether their perspective on medical terminology use leans toward prescriptive or descriptive;
- Discuss whether and how their work relates to issues of reuse and repositories; and
- For those addressing analytic contexts, to be clear about assumptions they make regarding:
    - data resources, acknowledging ways that these assumptions affect applicability of their findings to contexts with more or less access to semantic, aggregate, or patient-level empirical data;
    - study and data warehouse factors: single study, single database, CDMs, distributed or centralized research network data, vocabulary mappings.

4.7.2 Practical, managerial

Our account of code set practices and our synthetic work in classifying and organizing these practices into a comprehensible model can help code set developers adopt best practices. It will be helpful to situate any given project by the factors listed above. To characterize how these factors apply to a value set project, it should be useful to consult Tables 6, 7, and 8.

Our work can be applied to the development of structured workflows, forms, and checklists based on those factors. A form for collecting metadata and documentation for the code collection and code evaluation steps of the CEER process might look something like Table 9. For each

extensional code or intensional rule and expansion included or considered and rejected, the form provides fields for listing code or rule, vocabulary version, and expansion results. For each row, a source would be listed. Possible sources could be existing code sets, groupers, publications; a code set repository search for specific search strings; or navigation of vocabulary hierarchies starting from specific terms. Then the decision to accept or reject would be indicated along with reasons: inappropriateness of a code's meaning, the meanings of descendant codes, or aspects of how the code is used in relevant data capture workflows; the code's presence or absence in other code sets; zero or insignificant use in target databases; other empirical evidence or other arguments arising in discussion with clinical and terminology experts.

Table 9. Code-by-code metadata and documentation capture form for CEER code collection and evaluation steps with example of translated VSAC ICD-10 value set to OMOP standard codes

COLLECT CODES/RULES			EVALUATE CODES/RULES	
SOURCE	CODES		ACCEPT REJECT	REASONS
	INTENSIONAL RULE	VOCABULARY VERSION AND EXPANSION		
VSAC OID 123…, 2021-02-01	E11.* (Type 2 diabetes mellitus plus descendants)	ICD10-CM: E11.0, E11.1, …	Reject	Using the SNOMED codes these codes map to instead
OMOP `concept_relationship` ICD10-CM `Maps to` SNOMED	E11 maps to OMOP SNOMED `concept_id` plus descendants in OMOP	OMOP vocab 2024-03: 123, 456, …	Accept	In order to use OMOP standard codes. Have only mapped E11 to SNOMED Type 2 diabetes mellitus and got descendants of that. May want to check mappings of E11 descendants also.
User: E11.9 (T2DM without complications; has no descendants) `Maps to` OMOP `concept_id`	Exclude 4193704	OMOP vocab 2024-03: 4193704	Accept	Spot checking indicates that this code without out other T2DM codes is given as a rule out, for physical exam and testing, and may not indicate actual diabetes.

4.7.3 Technical

In order to make structured workflow and best practices practicable, the design of software, standards, and infrastructure to support code set development, curation, and reuse can benefit from CEER and the other classifications we offer here.

4.7.3.1 Code-by-code reuse

This opens a possibility for code-by-code *reuse*, that is, reuse of the effort and expertise put in and the knowledge created in the development of prior code sets *without requiring identification and reuse of a single existing code set.*

Calls for code set documentation accompany many discussions of code set quality and reuse, generally by asking the user for explanatory text about intentions, limitations, and the like. Motivating the sharing of digital knowledge or artifacts for reuse at all is challenging and efforts to do so often fail [148,149]. Burdening that process with additional documentation demands further complicates the problem. What we are proposing adds another level of difficulty as documentation of exploratory analysis is well-known to be challenging [150]. Yet, we believe that the basic processes of code set development are sufficiently structured and simple to allow software to automate much of the metadata collection and to make the steps requiring human input easier than open-ended text input prompts.

As this kind of software collects data like the form in Table 9, a database of how individual codes or intensional rules have been used or rejected for use in past code sets could provide code set developers guidance for their own process without requiring an understanding of the differences of all the alternative code sets already made for their phenomenon of interest. They could benefit from the work of the developers of all those past code sets without having to consider the individual code sets at all.

A key reason users decline to document their products for future use is doubt that anyone (including themselves) will ever benefit from their exertions. The structured, atomic commentary method proposed here reduces the effort required by eliminating the cognitive load involved in

reconstructing one's motivations later, or deciding on the frequency, timing, and detail level of notes. By standardizing note collection to the addition and deletion of individual codes, cognitive load will also be decreased for the potential re-user, providing them with information about specific codes so that reuse decisions do not require them to consider the quality and appropriateness of every entire code set they might reuse or be influenced by. This approach would also greatly reduce the amount of researcher-contributed documentation required to reach a degree of critical mass that makes the system worth using. And the value of contributions will be particularly high when they appear during future work of that researcher or their close colleagues.

A record of why codes were accepted or rejected for past code sets (whether for the phenomenon of interest or other phenomena) is likely to be directly helpful but could also tell future editors or re-users whether the absence in similar code sets of a code they think might be relevant is an oversight or intentional. It would even be useful to provide counterarguments for accept/reject decision so that later reviewers or re-users know that these arguments had been considered.

Beyond that, these notes do not just document the decisions made in developing a particular code set, they also document decisions made about that code across all the code sets or phenotypes where their use was considered.

4.7.3.2 Repositories; simple reuse is not recommended

We hope the analysis here has demonstrated that, insofar as RWD research requires empirical data for phenotype and code set development and evaluation, simple reuse of code sets is unlikely and inadvisable. The design features described above, however, do offer a promising

way for the knowledge embedded in prior code sets and phenotypes to be leveraged. This approach has much in common with [31], where value sets are broken into smaller units based on matching subsets of codes. Generalizing from that paper and this, we believe that the usefulness of value set repositories will emerge from techniques that extract codes, metadata, documentation from code sets; combine it with other information; and present it through interactive, exploratory user interface tools—rather than treating code sets as complete objects to consider for reuse as-is, with tweaking, or as digital objects that they can inspect and pull bits and pieces out of as needed.

4.7.4 Limitations and future work

In light of all we have learned since designing our survey, we would design a very different survey today. For instance, though we have been able to provide factors and reasons for code set development to be more or less rigorous, to take more or less time, a follow-up study should explicitly ask about the duration of code set development projects to learn how much time is or should be invested in code set development based on various contextual factors. We asked the number of studies performed per year but neglected to ask how many code sets respondents developed for a study or in the past x months.

We would have liked to give population estimates of the attitudes and practices we encountered. The results of the current study do lay groundwork for more systematic recruitment strategies, but identifying a representative sample of the universe of code set developers would still be challenging.

The field would greatly benefit from population estimates of formal code set validation practices, especially of the use of MRA-based gold standards and other kinds of reference standard.

Since this study was conducted, SG and HL have worked on N3C, supporting hundreds of studies. N3C users have developed thousands of code sets, which is problematic, of course, since there are now many code sets for most common conditions and covariates. To address this problem, N3C infrastructure support teams have worked on developing specific, N3C-recommended code sets for these. Prior to this experience, we had not encountered code sets being constructed in RWD settings without being meant for a single phenotype and study (though this does occur in the development of value sets for clinical quality measures and permissible value set uses). This paper describes why some of our participants are highly critical of efforts to develop analytic code sets for general use, but the need for such code sets is clearly felt in some settings and the field would benefit from guidance and caveats on their development.

We have recommended particular advances in software for code set management. The field would also benefit from a market scan of current tools, open and commercial, reviewing their support for the practices catalogued here. Our process model and classifications provide a groundwork for that kind of analysis.

Chapter 5. Narrowing the problem: Redundancy in value set repositories[19]

5.1 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. Value sets are groupings of these identifiers that facilitate data collection, representation, harmonization, and analysis. (We treat the term "value set" as more or less synonymous with "code set", "concept set", "code list", and "enumeration", which are also used in some contexts.)

This paper focuses on the use of value sets in the context of observational research using real-world data (RWD) [19]. Despite the use of hierarchical classifications and other data structures to signify concepts at different levels of granularity, value 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 algorithms for identifying patient records indicating O, C, T_1, and T_2. (Our use of the term electronic phenotype, or just phenotype, follows others in the field of observational research with RWD, e.g., [17,19,21,90], and can be confusing for those not

[19] This paper has been submitted to PLOS ONE as "Value sets and the problem of redundancy in value set repositories," by myself, Harold P. Lehmann, Lisa M. Schilling, and Wayne G. Lutters.. Preprinted on medRxiv [52].

accustomed to this usage. See Section 5.3.3 for a definition.) An essential step in such algorithms is to select patient records containing specified fields whose values are any of the codes in a value 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, value sets are usually the starting point for phenotype or cohort algorithms.

Crafting high-quality value sets is time-consuming and requires a range of clinical, terminological, and informatics expertise. Scholarly and practical efforts to address challenges in value set management (i.e., helping RWD researchers identify and select the set of codes best fitted to their hypothesis testing and analysis goals) [21,17,19,151,15,16,18,20,22,23] have resulted in value set definition and documentation standards [5,24–27] and in methods and tools for authoring value sets [31,152,32–34], for assessing value set semantics and quality [35–41], and for enabling and promoting value set sharing and reuse [28–30]. These papers demonstrate problems of bias and inaccuracy in value sets shared on public repositories and many present specific methods to improve value set development. Williams, et al. [17]—in a paper we used as a seed article for our literature review—performs a comparative review of the value set literature, offering nomenclature, a consolidated articulation of published knowledge on value sets, and a valuable catalog of recommendations for advancing technology for managing value sets.

The current paper offers a view of value 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 value sets, finding a diversity in real-world value set development practices and perspectives previously unexplored in the literature.

While there seems to be universal agreement on the importance of reusing value sets (or phenotype definitions containing value 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.

Value set repositories tend to contain many value 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 value 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 value sets for certain phenomena.

It has been implicitly assumed that value set repositories would improve and grow in utility as they gained wider and longer use. We ourselves have claimed that repositories would benefit by cooperating to consolidate or centralize in order to generate positive network effects by attracting wider audiences [18]. As we demonstrate, the opposite appears to be the case. With ongoing use, repositories accumulate redundant and low-quality value sets, making it increasingly difficult for a potential re-user to identify high-quality value sets appropriate to their needs. Positive network effects will only accrue if all contributions to a repository are dedicated either to improving existing value sets or making new ones when absolutely necessary.

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 value set development. These three dichotomies distinguish:

1. Permissible values versus analytic value sets. Permissible value sets are used in applications where data capture occurs (primary use). Analytic value sets are used in

analysis or research application (secondary use) in order to select records matching clinical conditions or events of interest.

2. Prescriptive versus descriptive perspectives on controlled medical vocabulary use. These tend to be held as implicit beliefs about coded concept, a prescriptive orientation is appropriate to permissible values contexts, while a descriptive orientation may be appropriate in secondary use, analytic contexts.

3. Semantic and empirical types of value set development and evaluation practices and the data they rely on. Semantic practices and data relate to vocabularies and meaning and are always necessary. A descriptive approach to identifying codes for an analytic value set, however, would require empirical analysis of patient-level data. Empirical analysis and validation are always desirable for analytic value sets, but it is frequently not feasible.

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

5.2 Methods

As noted, the intent of this research effort is to more deeply understand the diversity of real-world value 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 value 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[153] 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 value 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 value set authoring and reuse practices. The contextual nature of the interviews allowed them to demonstrate their tools and processes for developing value sets in person or via screen share.

The survey and interviews were approved by the University of Maryland IRB (#1405794-8). Recruiting began August 1, 2019 and ended on September 14, 2021. Taking the online Qualtrics survey required human subjects to read and sign our consent form. Interviewees signed a separate Qualtrics consent form. The survey and deidentified data are available at [154].

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 value set development and management efforts in a novel context. Their active participation in this wide range of projects has made them careful observers of value 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.

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).

5.3 Results and discussion

5.3.1 Diversity of value 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 10 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 value 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 10 and Table 12) and they worked on projects involving a range of vocabularies, domains, and data models (Table 11).

Table 10. Participant demographics and work contexts; study and/or value set development roles played by participant and other team members.

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	22
More than 5	13
Team size (people)	**#**
1	0
2 to 5	26
5 to 10	9
More than 10	1
Team size (organizations)	**#**
1	10
2 to 5	22
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

Of the nine most common tools our respondents reported using for value set development listed in Table 12, several are specifically designed for clinical or clinical research applications, providing support for authoring, sharing, and using value sets. The others are general programming and analysis tools with which value sets can be composed, evaluated, and used by linking to database resources containing vocabulary and patient information.

Table 12. Software tools, platforms, and repositories used in value 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 value sets.

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 11. Vocabularies, vocabulary domains, and data models targeted by participants' value 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	**VALUE 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		

5.3.2 Diversity of value set development practices

Value sets and processes for developing them vary in many critical ways. The effectiveness of a given value set development process and the accuracy of the value set it produces depend as much on the thoroughness with which methods are applied as on the selection of those methods. Table 13 lists the distinct tasks identified in the study.

Table 13. Value set-related tasks performed by survey respondents or their team members.

VALUE SET-RELATED TASKS	#
CREATE VALUE SETS	
Create value sets for local use	28
Create value sets for use by others	27
Vocabulary search	15
Vocabulary navigation	16
Consult clinical or terminology experts	16
Optimize value set to more parsimonious expression by replacing codes with intensional rules where possible.	5
Translate value sets across terminologies	21
Add value sets to repositories	24
USE EXISTING VALUE SETS	
Use value sets created by others	30
Use value sets from repositories	20
Use value sets from publications	8
Manual or automated comparison with existing value sets	29
EVALUATION	
Evaluate value 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/value set result counts may be taken by developers to mean the value 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 value set—preferably by executing its containing phenotype algorithm. (Requires access to patient-level detail to generate value set counts.)	13
Sensitivity analysis of changes in results caused by modification of the value set.	4
Test value 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 value sets	22

A particularly important factor shaping value set development practice is whether the value 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 [15,16,25–27,31,32,35–40] asserts the importance of reuse in addressing problems with value set quality. We suspected

at the outset of this study that reuse was uncommon, as there is considerable evidence [18] that reuse is fraught with difficulties and that repositories accumulate many value sets ostensibly representing the same clinical phenomenon. Our field data [154] show 30 (83%) of our respondents reuse value sets made by others and 20 (55%) use repositories to find value sets for reuse. Many participants mentioned sharing value sets to public or private repositories—about a third to the Observational Health Data Sciences and Informatics (OHDSI) ATLAS web interface [116], a third to VSAC, and several to other repositories, publications, and research networks.

5.3.3 Permissible values versus analytic value sets

We distinguish two general types of value 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 value 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. (In other contexts, such as data harmonization and clinical quality measures, value sets are used in more ambiguous ways that have both permissible and analytic qualities.) There are other use cases for value sets but the differences between these two contexts (data capture and RWD analysis) will show why the distinction is needed.

The distinction is not about the digital structure of value 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 value 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

Table 14. Contexts for value set development.

CONTEXT	REUSE REPOSITORIES	PERMISSIBLE / ANALYTIC	PRESCRIPTIVE / DESCRIPTIVE	SEMANTIC / EMPIRICAL EVALUATION
Value sets for data capture	VSAC	Permissible	Prescriptive	Semantic
Other value sets for terminology services (e.g., FHIR, CTS2)	VSAC	Permissible	Prescriptive	Semantic
Clinical quality measures	VSAC	Both	Both	Mostly semantic
Single study, single database		Analytic	Descriptive	Both—but empirical is vital and possible
Network study or multiple related studies	ATLAS, N3C	Analytic	Descriptive	Both
For analytic reuse but not for a specific study, database, or question	N3C	Analytic	Descriptive	Mostly semantic

[35,36,38,39,29,44,105,139,140,28,142,83,143,144,106]. 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 value sets necessitate different methods and tools to author, validate, share, and reuse value sets. Table 14 categorizes value set development contexts using this dichotomy and two others described in the next two sections.

Figure 3 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 for a diabetes complications data element—to populate dropdowns and constrain data element values.

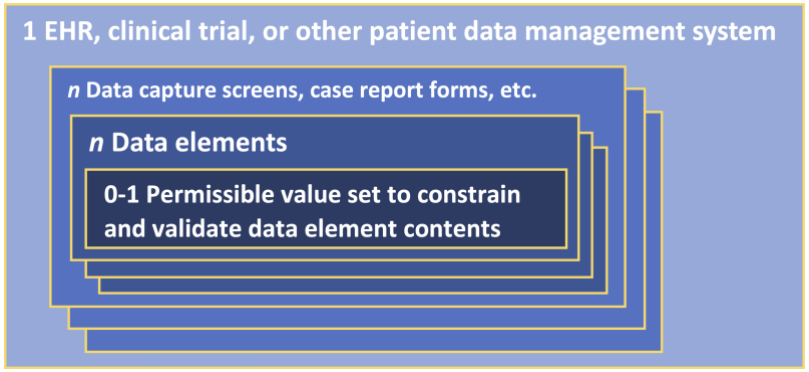

Figure 3. Permissible value sets in context.

In RWD studies, analytic value 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 [21]. 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 (e.g., narrative notes, images, EKG or other device output), but insofar as terminology codes are used in the algorithm, a phenotype will include one or more value sets as diagrammed in Figure 4 and may use temporal and conditional logic in performing set operations on the groups of patient

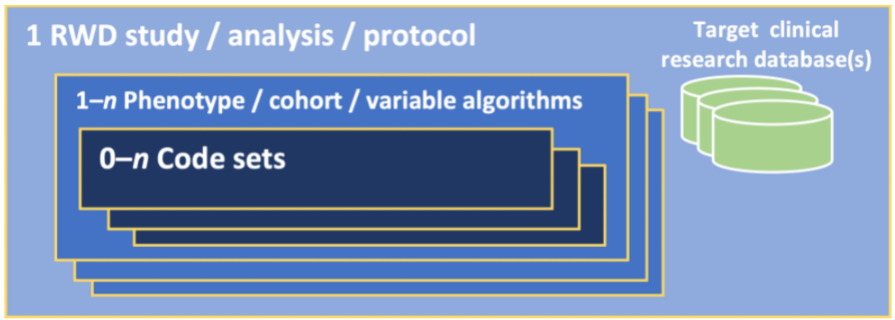

Figure 4. Analytic value sets in context.

records matched by different value sets. However, phenotypes can also be as simple as a single value set, the algorithm consisting of nothing but the selection of patient records containing one of the codes in that set. (See, e.g., "Finding Existing Phenotype Definitions" in the phenotyping chapter of the online textbook, Rethinking Clinical Trials [90] which lists value set repositories alongside repositories of more complex algorithms as sources of reusable electronic phenotypes.)

Permissible value sets are generally developed for specific clinical data systems, often for a single institution. Analytic value 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.

5.3.4 Prescriptive and descriptive perspectives on value sets

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

- On permissible value sets for permissible values or clinical quality measures (CQM) [35–39,28,29,106,44,105,139,142,83,143,144,140]; 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;
- On phenotypes (cohort selection) for RWD research [21,24,90,146,147]; and
- On analytic value sets (often called code sets) for use in phenotyping applications [17,18,30,145,152].

The computable phenotype literature sometimes conflates phenotypes and analytic value sets. Where it does discuss reuse, the focus is on phenotype rather than value set reuse. While some in

this group do not believe that sharing value sets separately from phenotypes is worthwhile, we have seen that reuse does occur and there is demand for value sets that can be used across phenotypes. The permissible value set and value set literatures both tend to focus on value set repositories and reuse.

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

Whether one considers expert authority or empirical analysis as the primary means of value set evaluation can reflect almost ideological beliefs about the nature of value sets and medical 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 [155], the imposition of prescriptive terminology is, of course, the foundational

purpose of standardized medical vocabularies and arguably the foundational practice of medical informatics generally [47,156–158].

Exemplifying a prescriptivist orientation, Winnenburg, et al. 2013 suggests that a value 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 value 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 set**s and data capture contexts; and descriptivist for analytic value sets and RWD research.

5.3.5 Semantic versus empirical methods and resources

Williams 2017 alludes to a central tension in the choice of methods for value set validation [17]. 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.

5.3.5.1 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 value sets from "previous[ly] published results" (P01), "examin[ing] the literature for validation studies" (P41), or "validated codesets when possible" (P04). (Value sets are available from value 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 value set developer; and groupers such as Clinical Classifications Software Refined (CCSR) [59,60].

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).

5.3.5.2 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 value 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 value set development and validation. Choice of evaluation methods can be guided by clinical nuances of the research question or how the value 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 set**s. (P34)

To unpack that statement a bit, a highly *sensitive* value 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 value set is needed, comparison with existing value sets could help to make sure that appropriate codes are not missed. A highly *specific* value 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 value set/phenotype modifications will help.

Evaluating the accuracy of a sensitive value set requires a thorough semantic exploration to identify all codes that could indicate the condition of interest, while evaluating the accuracy of a specific value set should involve empirical examination of matching patient records to prevent false positives. (Software tools are available particularly to help value 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.)

Table 5 lists general contexts in which value sets are used and relates them to the three literature categories listed above, to value 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.

5.3.6 A taxonomy of reasons for value sets to differ

While repositories make it possible to share and reuse value 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 value sets for (ostensibly) the same clinical concept may differ into three categories: semantic, empirical, and due to error. These are summarized in Table 15.

Table 15. Reasons for value sets with same topic to differ in definition and composition.

<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 or database 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>

5.3.6.1 Semantic reasons for value sets to differ

Two value 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 value set for a cardiology study may require a different set of codes than an endocrinology study.

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

Vocabulary issues. Value 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. (E.g., CDMs like OMOP may require translation or harmonization of codes in patient records to agreed-upon vocabularies.) 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 value set composition. (This not strictly a semantic issue but a fact of value set use.)

- A value set designed to target a diagnostic condition might use evidence from other domains of clinical data, e.g., drugs, procedures, lab tests. (It is recognized in the literature that phenotypes benefit from the use of multi-modal, multi-domain data [61].)
- A value set may be designed in the knowledge that it will produce false positives or negatives if these will be corrected by logic or other value sets at the phenotype.

5.3.6.2 Empirical reasons for value 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.

Regional, institutional, or clinical specialization coding practices. A single meaning may be expressed using different codes in different places. (This possibility was mentioned by our participants and seems to be a relatively common belief, but we have not encountered specific examples.)

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, value sets for a given phenomenon can differ without giving potential re-users any basis for choosing between them.

5.3.6.3 Errors

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

Codes mistakenly left out or mistakenly included. When value 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 value 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 value set or its containing phenotype may show perfect sensitivity (low false negative rate) and specificity (low false positive rate) while differing from a value set based on another gold standard.

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

5.4 Conclusion: Leveraging and mitigating redundancy in value set repositories

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

> [Reusable value 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 set**s, 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 value sets or indications of a value set's having been vetted and validated, what we see in repositories is proliferation and clutter: new value 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 value sets for a given condition often leads value 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 value 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 value set's intention and provenance.

There is a tension regarding how many value 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 value sets. On the other hand, tight coherence with the research question—"fitness for use"—argues for

customizing a unique value 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 value set creation, we must be able to understand when an additional value set for a target condition may be needed or not. The taxonomy in Section 3.6 may help in reconciling differences when multiple value 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 value sets or codes.

When a new researcher creates their own value 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 value 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 value 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 value set developers through the process. Toward that end, we offer the following recommendations.

5.4.1 Advanced, automated comparison tools

In our professional experience we have seen instances where trust in what were considered authoritative value sets broke down when comparing them to other value sets. One participant, P16, performed an automated comparison of many alternative value sets for depression, using

the differences and similarities to create a trustable value set without having to trust any of the input value sets individually.

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

5.4.2 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 value set level; it could be enormously helpful to collect metadata to capture value set developers' reasoning for including or rejecting specific codes. (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 value set, but this feature has not yet been developed to the point of being useful—nothing is currently done with users' input.) A combination of automated process data capture and timely, minimally obtrusive user prompts could provide code-level metadata that could be displayed as future value 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 value set, record a reference to that value 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.

5.4.3 Expert or automated curation

Terminology experts on the N3C infrastructure staff have developed "N3C Recommended" value sets for commonly studied topics (conditions, medications, medication classes, measurements, procedures). VS-Hub was specifically designed to facilitate that endeavor. VS-Hub, so far, has only attempted to make it easier to cull the best out of available value sets for a given condition or event, it has not attempted to force users to review relevant value sets and either improve one of those or make sure that a new one is genuinely needed. If we, as a field, hope to see value set repositories increase rather than decrease in quality and usefulness over time and widening use, strong curation will be necessary to exclude redundant, unfinished, or otherwise low-quality value sets. Such curation could be done by humans, software, or both.

The requirements and recommendations in prior literature have not been sufficient to guide the design of software that could make effective leveraging of shared value 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.

Chapter 6. Implementing solutions: VS-Hub, software for developing and curating high-quality value sets[20]

6.1 Introduction

The importance and productivity of observational, *in silico* research based on electronic health records, reimbursement claims, and other real-world data (RWD) has exploded over the past ten to fifteen years. Thousands of researchers in networks like OHDSI, PCORNet, All of Us, and N3C[21] are able to leverage vast, multi-site data sources harmonized to common data models (CDM) using open-source software and infrastructure to perform replicable, FAIR research at hitherto unheard-of speed and low cost.

A particular problem area in the execution of RWD studies is in the development of analytic value sets: groups of controlled medical terminology codes used to query patient records in the computation of cohort or phenotype membership and study variables.[52][22] Designing the algorithms used in executing research or safety studies (to compare outcomes for alternative treatments, for example) requires an understanding of observational, retrospective study design; the clinical topic of interest; and the care settings and operational workflows shaping the data available for study. Formulating the conditional and temporal logic for the overall study and specific cohorts and variables entails unavoidable thought and work. The selection of codes for

[20] This paper was submitted for the 2024 AMIA Annual Symposium. It was returned with comments and will be revised and submitted elsewhere.

[21] Observational Health and Data Science (OHDSI),[7,8] The National Patient-Centered Clinical Research Network (PCORNet),[50] All of Us[45], and The National COVID Cohort Collaborative (N3C)[10]

[22] We refer to defined sets of controlled medical vocabulary codes as value sets, though the N3C and OHDSI communities call them concept sets, the literature specifically discussing them in the context of RWD research generally calls them code sets, and some ontology communities call them enumerations. In other contexts, they may also be called code lists, groupers, or term sets.

the value sets used in these algorithms seldom receives as much attention despite the fact that these value sets determine the selection of patient data that serve as input to the algorithms.

Creating value sets can be easy, for instance, by doing a string search for vocabulary terms and then, perhaps, navigating around vocabulary hierarchies to find relevant related terms. This might result in a perfectly adequate value set, or not. It is often not feasible to perform a thorough empirical validation based on a gold standard of patient records marked as having or not having some particular condition or phenotype. Lacking that, the quality of a value set is not directly measurable; it is indirectly inferred by diligently applying the best practices in its creation: thorough reviews by clinical and terminology experts, cross-referencing with similar value sets, review of matching record counts, and spot checking of those records.[20]

As a field, we know these best practices and see occasional papers exhorting us to use them or offering improvements to one or another, yet quality problems persist.[18,21,128] Particular attention has been given to the sharing and reuse of value sets[17] in public value set repositories like VSAC[29] and Clinical Codes[30] or repositories integrated into larger RWD research platforms like OHDSI/ATLAS[7] and the N3C Enclave.[10]

Recognizing the effort and skill required to craft high-quality value sets, the repository developers offer tools to encourage sharing and reuse, hoping that researchers will take advantage of others' work, building on it where possible instead of repeating it. What we see in actual repositories, however, is a proliferation of value sets for common features of interest (diagnoses, treatments, etc.) [52] Those needing value sets either do not think to check for existing value sets and create their own or, finding many candidates for reuse with no easy way to discern their quality of appropriateness for their needs, they again make their own.

In the N3C community, we and our colleagues have made concerted efforts to encourage reuse by asking value set authors to provide metadata about the provenance, intentions, and limitations of their value sets and by providing extensive documentation and training to promote best practices and reuse. These efforts have not led to discernable improvement. Because current tools can make it time-consuming and difficult to follow best practices—e.g., expert review of large value sets when concepts are not presented hierarchically; lack of effective interfaces for comparing candidates for reuse; lack of or inconvenient access to term usage counts—much of the work required to make a high-quality value set may not happen.

We have come to believe that the only way to get users to reuse appropriate value sets or, when creating their own, to follow best practices, dedicating sufficient and fitting effort to create them well and prepare them for reuse by others, is to provide software that pushes them to do these things, mostly by making it easy and obviously beneficial to do them.

We present Value Set Hub (VS-Hub)[23] as a platform for browsing, comparing, analyzing, and authoring value sets—a tool in which the presence of multiple, sometimes redundant, value sets for the same condition strengthens rather than stymies efforts to build on the work of prior value set developers. VS-Hub introduces several innovations to the state of the art for value set authoring platforms.

In the Design section below, we describe the goals and requirements that have driven VS-Hub design and the tools used to build it. The implementation section provides an abbreviated account of the development trajectory, the evolving needs and priorities that have driven

[23] The software has been called TermHub and that name lingers in various places; we changed it to VS-Hub to avoid possible trademark infringement.

implementation of specific features, a sense of the overall architecture, and description of the features on the two primary user interface (UI) screens. The evaluation section gives quantitative and narrative description of VS-Hub's actual use in the development and maintenance of value sets over the past several months. The Discussion section addresses generalizability and opportunities for further work.

6.2 Design

VS-Hub's developers work as part of a team whose mandate includes the curation, development, and maintenance of recommended value sets for conditions and electronic phenotypes (i.e., cohort selection or research variable algorithms) commonly needed for RWD research. Understanding that each research project is unique and that researchers will sometimes require more than a one-size-fits-all value set, we have developed VS-Hub both to serve our own office (and other informaticists with terminology expertise who similarly endeavor to build or curate value sets for use beyond a specific project) and to serve the more general audience of those looking to find or create value sets for a specific need.

Specifically, the software should:

- Maximize the information immediately visible or rapidly available to support user decision-making as they review existing value sets and reuse or revise them for their own studies;
- Encourage the user to find and review existing value sets most relevant to their topic before creating a new one, showing them summary data regarding value sets of possible interest; counts of definition and expansion concepts; matching patient and record counts; author, version, intention, provenance, and other metadata;
- Make users aware of the semantic neighborhood of the concepts they are considering by showing (visualizing) value set member concepts in the context of their vocabulary hierarchies and other semantic information when available (e.g., concept domains, mappings, membership in other value sets);
- Provide any available empirical information about individual concepts, such as patient, record, and descendant record counts or validation data;

- Present many-to-many comparison of selected value sets;
- Highlight differences between selected value sets;
- Encourage the development of parsimonious value sets—that is, value sets defined intensionally using a minimal set of high-level concepts that will be expanded to include or exclude their descendants.[24]
- Encourage thoughtful review and maintenance of value sets after vocabulary updates, showing concepts added or removed when expanding definition (intensional) concepts using current vocabulary versions;
- Effectively hide data when value sets include too many concepts for performant display in browsers or comprehension by users (e.g., by collapsing very deep or very wide descendant trees), providing clear summary of hidden information to facilitate discovery and display.

The long-term aim of VS-Hub is to serve as a central value set exchange and authoring platform, interoperable and synchronized with external sources of value sets such as VSAC, N3C, ATLAS instances, and FHIR value set resources. Though it is designed for generalizability, implemented features so far have been tailored to the evolving needs of its initial audience, the thousands of researchers in the N3C community. This community conducts research using the N3C Enclave, a secure environment built and hosted with Palantir Foundry for managing and analyzing harmonized, multisite data for 22 million COVID patients and controls. Its data structures follow the OMOP CDM and it uses the OMOP vocabulary system to harmonize and integrate concepts from many source vocabularies (e.g., SNOMED, RxNorm, LOINC, CPT, ICD10CM).

After extensive work with Palantir engineers building the Enclave's Concept Set Browser and Editor, it became clear that many of the aims listed above would not be effectively implemented because 1) engineer time for the project was limited, and 2) the Palantir Foundry UI development

[24] This is generally advantageous because 1) it makes it easier to understand the intent of the value set; and 2) if future vocabulary updates add descendants to an intensionally defined code, they will be included in expansion without needing changes in the value set definition.

tools were not designed for the kind of information-dense display and rapid interactivity we believed necessary. Hence, we determined to build VS-Hub outside the Enclave using standard web development tools. Additional motivations for that choice include being able: to (eventually) accommodate and translate between many value set repositories and frameworks; to allow value set review by subject matter experts who don't have Enclave accounts and generally serve the wider informatics and research communities; to facilitate rapid feature development using our tools of choice; and to allow and invite open-source contributions from informaticists and software engineers beyond our team.

Software tools and platforms used in the implementation of VS-Hub are listed in Table 16.

Table 16. Software tools and platforms used.

TOOL	TYPE	PURPOSE
PostgreSQL	Database server	Backend data management
Python	Language	Backend server, external system synchronization, unit testing
FastAPI	Python package	Backend server
OAK [160]	Python package	Semantic graph query for vocabulary data
NetworkX	Python package	Semantic graph query for vocabulary data
JavaScript	Language	Frontend web interface
React	JavaScript package	Frontend user interface
Graphology	JavaScript package	Semantic graph query for vocabulary and value set data
Jest	JavaScript package	Frontend unit testing
Playwright	JavaScript package	Frontend end-to-end testing
Azure	App hosting service	Hosting of database and backend and frontend applications
GitHub	Code hosting service	Version control; management of Azure deployment and continuous integration testing

6.3 Implementation

Though the bulleted aims listed above have all been achieved to some degree, we have had and continue to have a long list of planned features to implement them more effectively. Design

and development have often been driven by specific projects our team has been responsible for and features have been implemented as resources allow. We will give a brief account of the development trajectory.

Our first attempt to provide the community with a library of N3C-recommended value sets was based on importing credible value sets from VSAC. Toward that end, we built a framework for storing value sets, using VSAC APIs for import and Palantir Foundry APIs for upload to the Enclave. This strategy also required conversion from source vocabulary codes to OMOP standard concept IDs. This was straightforward for conversion from vocabularies that OMOP counts as standard (SNOMED, RxNorm, etc.); but from other vocabularies (like ICD10), mapping to standard proved problematic, especially due to pre/post coordination issues.[136]

As these and other issues affected the majority of the value sets we needed, we turned to using and improving value sets already in the Enclave. There we faced a problem of massive redundancy and clutter. For instance, of the 5,000 value sets in the Enclave (7,600 if counting versions), 80 contain the word ‘diabetes’ (260 if counting versions).

We extended our framework to download and update value sets from N3C and we built a user interface for browsing and selecting from these (including display of relevant metadata), recommending related value sets, comparing those selected to each other, and presenting member concepts with an indented tree based on vocabulary hierarchies. We have struggled and tried many different approaches to dealing with the problems of recommending related value sets, displaying several value sets at a time, and retrieving and displaying the amounts of data involved when handling larger value sets.

VS-Hub's search, browse, recommend, and select screen (Figure 5) treats every selected value set equally, so the related value table presents a list of every value set sharing one or more concepts with any of the value sets selected so far. With the union of all the concepts belonging to the selected value sets as a basis, the related list shows precision and recall figures and other counts, allowing the user to sort on these columns to find related value sets most relevant to their needs. In order to calculate the shared concepts, precision, and recall columns for the three value

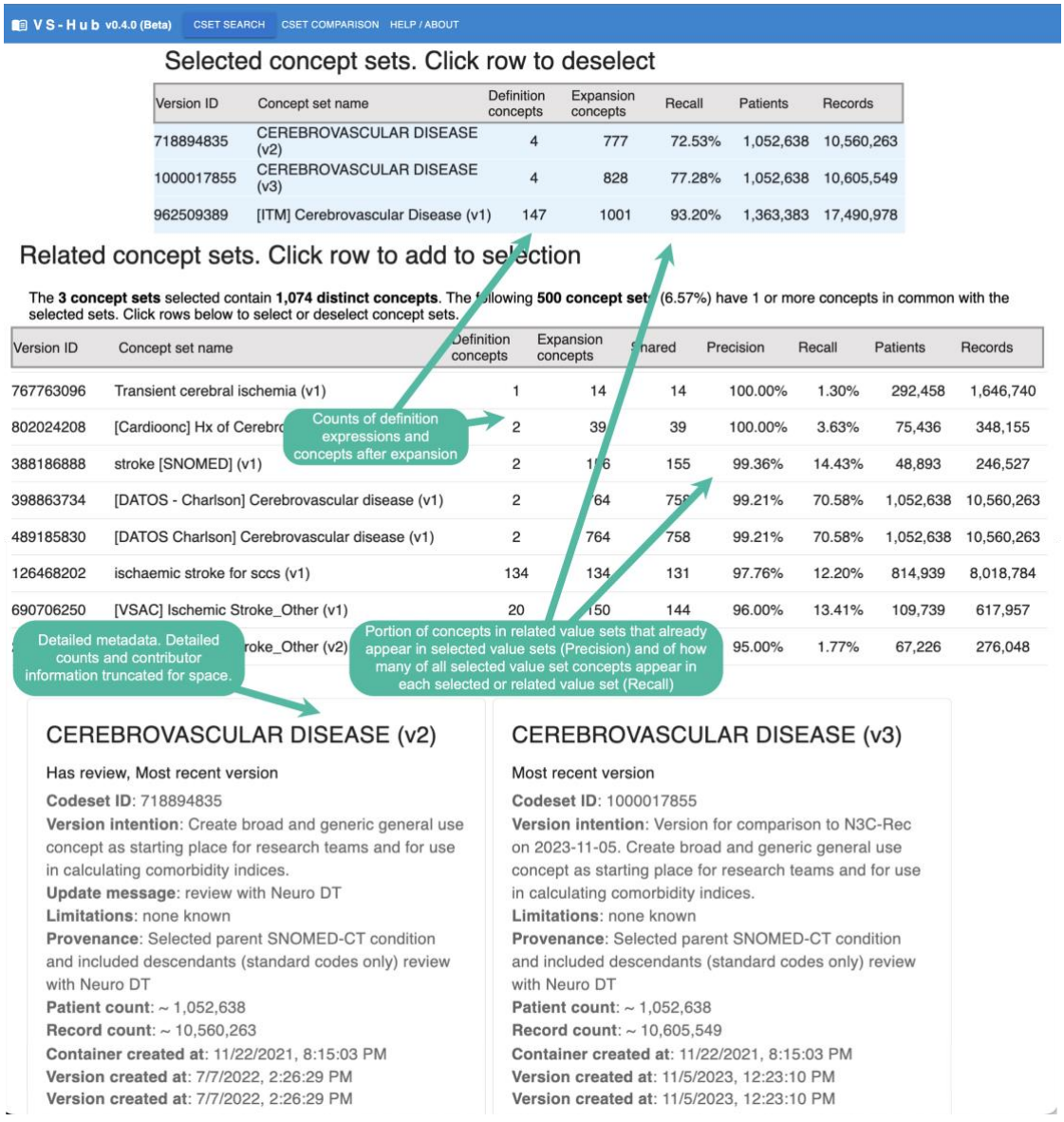

Figure 5. VS-Hub's search, browse, recommend, select screen.

sets selected in screenshot in Figure 5, we take their total 1,074 distinct member concepts, retrieve the 500 related value sets that contain at least one of those, then retrieve the members of each related value set: 1,225,496 total, 304,472 distinct concepts.

We give these numbers not to be tedious but to convey that even for moderately sized value sets (100 – 1,000 concepts), the calculation of these figures can involve substantial data processing. In order to keep the application from being painfully slow, we have tried a variety of optimization and caching strategies (discussed below in Lessons learned).

VS-Hub's display, comparison, and authoring page (Figure 6) presents a table of concepts (first column of each row), metadata (middle columns), and value set membership (rightmost columns). Another column on the right would appear for constructing a new value set. We omitted that and description of its UI features for reasons of space.

The OMOP vocabulary system used by N3C and many of its source vocabularies are structured as polyhierarchies or directed acyclic graphs (DAG), that is, pairs of concepts (terms, codes) are connected by directed edges, relating them as parent/child or source/target, such that a parent generally has many children and a child sometimes has more than one parent.[25] Intuitively users think of groups of related concepts as forming a tree, like a file system directory tree. VS-Hub follows the convention of representing such trees as collapsible, nested (indented) lists, though we have had to re-implement the indented list to deal with several complexities.[26]

[25] Though concepts have many types of relationships, VS-Hub currently displays only is-a/subsumes and other hierarchical relationships captured in the OMOP `concept_ancestor` table.

[26] We are also working on a node-link diagram that will more intuitively convey the DAG structure.

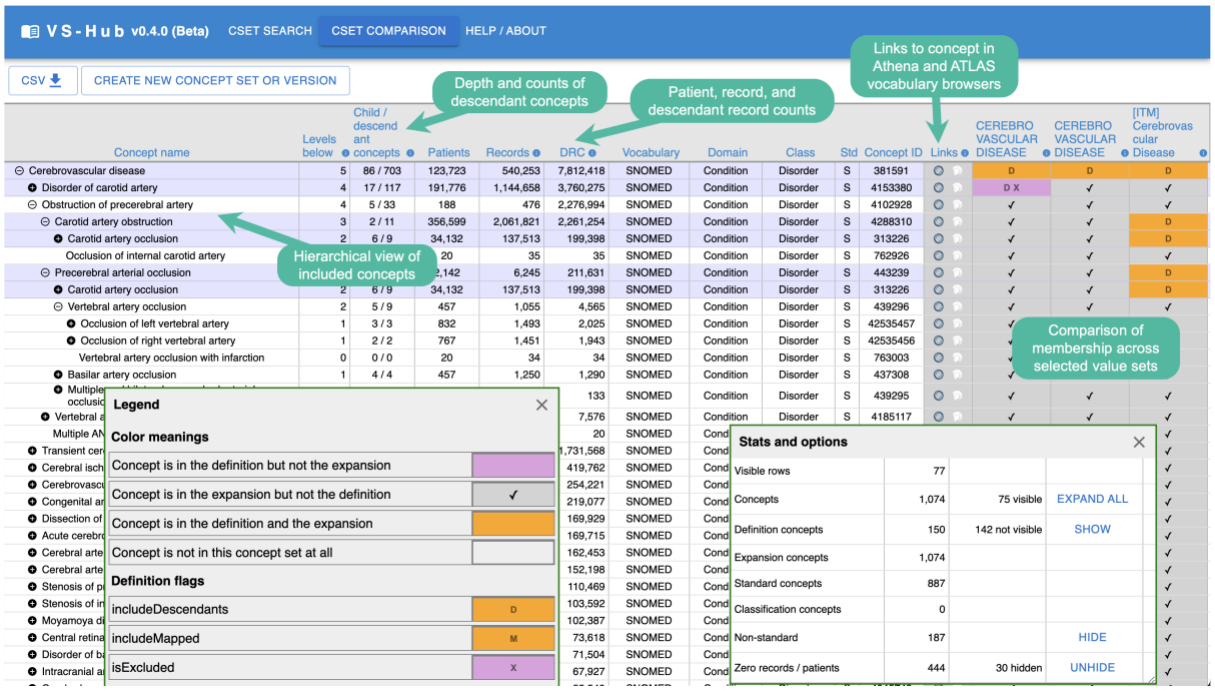

Figure 6. VS-Hub's display, comparison, and authoring page

We use available React components for frontend UI elements where possible. Available nested list controls only allow individual items to be displayed as a single string or React component. Given the amount of information we want to convey about each concept, a tabular display was necessary. We represent nesting by indenting the first column (concept name) of each row and showing an expand/collapse (+/–) icon on rows with children. One shortcoming of our current implementation using react-data-table is that we are only able to add or remove rows by re-rendering the whole table. It would be better to animate the expand/collapse operations.

In order to construct the nested tree display, we select a subgraph of the full OMOP vocabulary DAG[27] consisting of all the concepts included in the selected value sets and recurse through it, showing concepts indented by level, multiple times if they have multiple parents, and

[27] We have this as a NetworkX directed graph generated from all rows of the `concept_ancestor` table where `min_levels_of_separation = 1`.

sorting each level by descendant record count or other columns of the user's choice. Because some value sets are so large that displaying all their concepts will crash the browser tab, the hierarchy is initially displayed with all nestings collapsed, allowing the user to expand individual rows or click Expand All from the Stats and options dialogue. As seen in Figure 6, concepts that appear in the definition of at least one of the selected value sets are shaded in light purple. As indicated in the Legend, concepts that appear in the expansion but not the definition of a value set display a checkmark at the intersection of the concept and that value set. Intensional definition concepts show the options defined for their expansion (D for including descendants, M for including mapped concepts, X for excluding from the expansion—see cells in top right of Figure 6) and are shaded orange or purple depending on whether they appear in the expansion.

A problem we have had to address in the hierarchical display is that sometimes the set of concepts will include pairs that do have an ancestor-descendant relationship but not the intervening concepts that connect them. In that case, the generated subgraph will not contain an edge connecting them and the descendant will act as a root of a disconnected component. Since users want to be aware of relevant ancestor-descendant relationships, we fill in the missing nodes. Getting this right was a trial involving many (mostly misleading) conversations with ChatGPT. Eventually, with much care, we developed a unit test that handles the various edge cases we found in different value sets, diagrammed in Figure 7. The final algorithm works by traversing the full graph upward from each subgraph leaf node, capturing ancestor nodes up to any whole graph root, and discarding any of these that do not appear between two subgraph nodes.

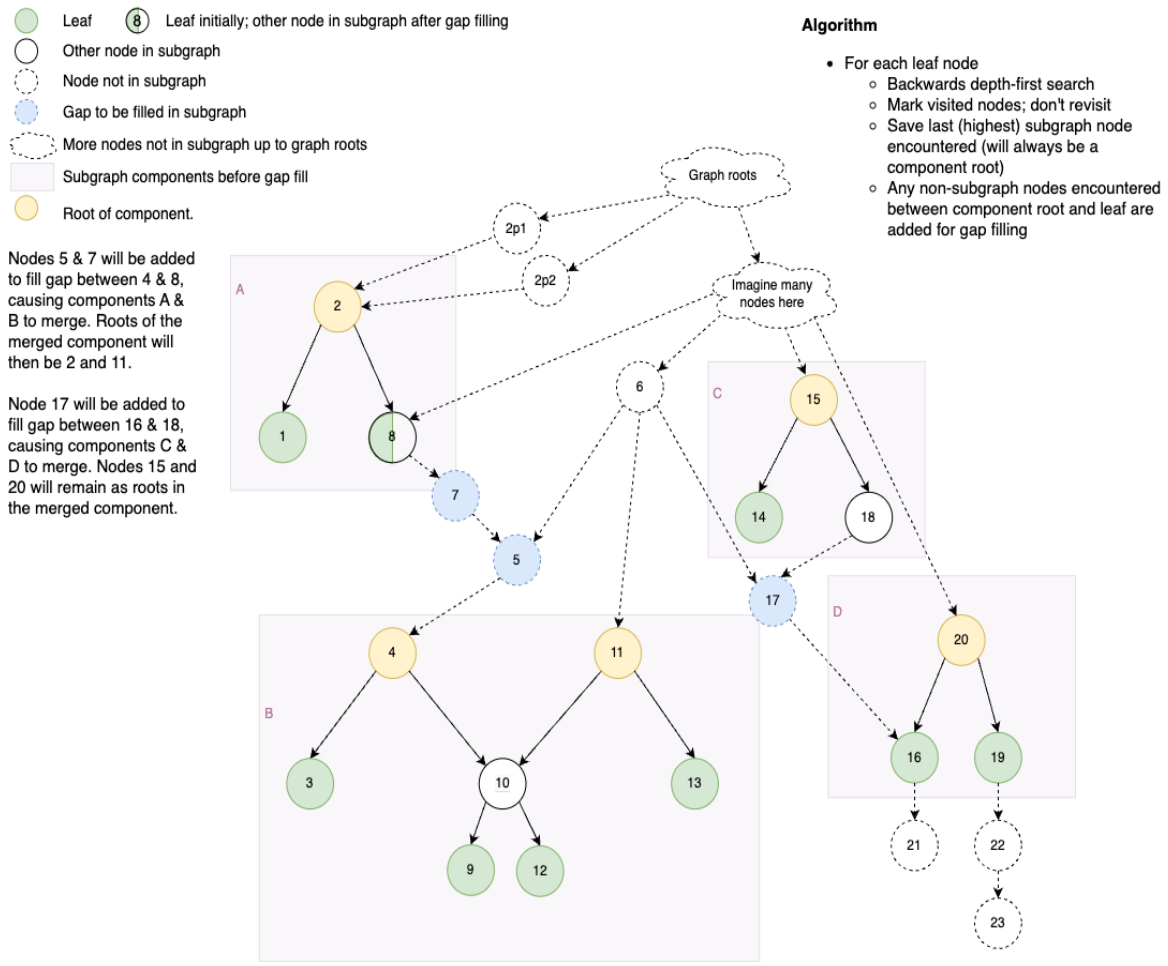

Figure 7. Diagram of gap-filling algorithm

6.4 Evaluation

Between November 2023 and March 2024, we collected backend server usage logs. Since we use caching to avoid redundant server calls, these logs do not capture analysis of already-downloaded data. After removing 4,999 log entries of testing or use by VS-Hub developers, the remaining 23,704 records represent use by our target audiences. Testing was agile and incremental amidst pilot deployment as beta software. Some of the 23,704 records analyzed include users just trying the software out rather than performing a specific task.) Table 17

provides summary data captured by usage logging. These figures make clear that VS-Hub is being used beyond the team that is building it.

Table 17. Usage and application statistics

Measure	Value	Notes
Total log records	28,703	All API calls that invoke tracking
Log records -- developer IPs removed	23,704	
Log sessions	12,014	Calls to multiple API endpoints from the same browser page are grouped together as “log sessions”
IP addresses	253	
User API call errors	11	
Developer API call errors	183	

Figure 8 shows a multi-modal distribution of VS-Hub usage levels; that is, over the five months of log capture, about 80 users made fewer than five page visits; about 70 made a few more visits; about 50 made between 15 and 20, and 10 or so made around 40 visits. (This chart does not include use by developers.)

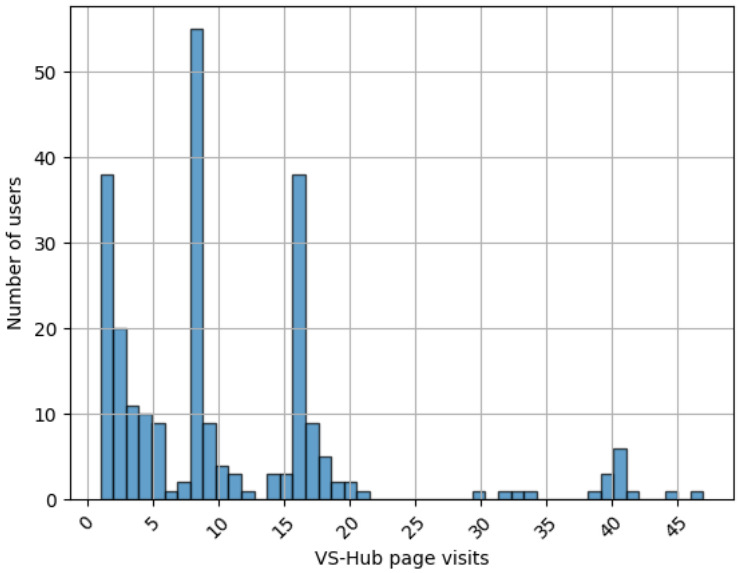

Figure 8. VS-Hub use distribution

Table 18 shows the range of value set sizes in the N3C repository as a whole and the count of concepts (from one or more value sets) returned from VS-Hub API calls that return concepts.

Table 18. Concepts in value sets and VS-Hub calls

Value set size across N3C	
Concepts	Value sets
1 - 9	774
10 - 99	1,862
100 - 999	2,596
1,000 - 9,999	1,357
10,000 - 99,999	537
100,000 - 999,999	64

Concept counts in VS-Hub calls	
Concepts	VS-Hub calls
1 - 9	15
10 - 99	1,326
100 - 999	620
1,000 - 9,999	129
10,000 - 99,999	50
100,000 - 999,999	24

6.4.1 Discussion

6.4.2 Lessons learned

The hardest problem we faced and where we made the most mistakes was in generating the indented tree display for large value sets. We should have addressed it from the first rather than waiting until code had already been written and users started needing to work with value sets larger than the application could handle. Even then, our initial approaches were overly complex and only appeared to work correctly (because we had not formulated good tests): we generated the indented tree (with duplicate rows for concepts with multiple parents) at the server, hiding subtrees after a certain depth or where direct children of a given node exceeded a certain number. Through trial and error we found that limiting concept graph download to edge lists and concept metadata worked, if a little sluggishly, even for our largest value sets. From there, we were a little surprised to also discover, performing graph operations using graphology.js to generate indented rows and calculate descendant records for collapsed descendants was very fast.

We had long wondered why other value set browsing/editing tools didn't directly display selected concepts in the context of their vocabulary hierarchies; it seems such an obvious and beneficial feature. The answer is, apparently, that it's hard. Value sets range vastly in size and graph topology (width, depth, connectedness, polyhierarchy); designing a user interface that handles all cases well is a serious challenge. This has proven to be an historical development

challenge, with few implementations to our knowledge. Nevertheless, we persevered to the point of proving that it can be done and have built an interface that does a credible job of it, with plenty of room for improvement of course.

As to why previous tools hadn't used a tabular presentation to compare concept membership across multiple value sets, getting it to work required creative use of React, which is a more flexible UI framework than those used by other value set authoring tools, and this display makes the need for hierarchical concept presentation more painfully clear.

An application of this scale should not be built by one and a third programmers but by a software development team including project management; dev ops; quality assurance to implement unit, end-to-end, and user testing. A lot of mistakes and blind alleys could have been avoided, especially if we had been able to begin with a test-driven development approach. On the other hand, the application has provided significant value to our group and to the N3C community at large, and we built it with the resources we had.

Frontend caching was essential; sluggish initial load times for sizable value sets would be tolerated, but when every page reload or change in the list of selected value sets took just as long, users grew frustrated. An early approach used a library that cached each distinct API call (accounting for differences in parameters) but wasn't sufficient. We built an overly complex but functional system for caching results granularly so that subsequent calls for overlapping data could use the cache and only download items that had not been retrieved. That is, if a related value set is added, it is only necessary to retrieve metadata for concepts not contained in the value sets already selected. We encountered many problems caused by cached data that should have been cleared for various reasons. We tried a variety of solutions including, horribly, asking

users to hit an Empty cache button on the help page whenever the application misbehaved in case bad cache data was the culprit. Our current imperfect strategy is simply to automatically clear the user's cache every 24 hours. Clearing least recently used cache data would have been helpful, but the complexity of our granular caching approach makes it difficult to implement.

6.4.3 Limitations and future work

VS-Hub needs additional features for users to explore vocabularies for candidate concepts by string search or by exposing concepts related to those displayed but not currently included in the subgraph.

Hierarchical concept display makes it considerably easier to author parsimonious intensional value sets, which we consider a best practice, but it needs to go further and help users identify common parent or ancestor concepts when a more parsimonious definition is possible. The approaches we have tried so far have compromised usability by adding excessive polyhierarchy or bringing in unwanted descendants.

When vocabulary changes lead to changes in value set expansion, VS-Hub makes it easy to see these changes in context, but users want to understand, especially, why certain concepts no longer appear in the expansion and help finding replacements if appropriate. We have not yet tried to address this need.

The current implementation has high memory demands.

We envision VS-Hub being generalized and used more widely by 1) accommodating and storing data from more value set formats (e.g., VSAC, FHIR); 2) connecting it to external value set repositories; 3) allowing users to build and store value sets optionally synced to theN3C

Enclave and other repositories; 4) allowing term usage counts and value set patient/record counts from multiple sources; and 5) by allowing institutions to host their own VS-Hub instances, connected to whatever private or public value set repositories, analysis platforms, and data sources they like. These extensions will make VS-Hub useful to a very wide community of RWD researchers and analysts. We do not currently have resources to implement them but hope to attract open-source software developers who could help.

VS-Hub represents a significant advance in the technology available for analyzing and authoring value sets. It removes obstacles that value set developers face in following best practices and making use of prior work. We have demonstrated the importance of the features it introduces, reviewed challenges encountered in implementing them, and provided lessons learned to ease the path of others who may attempt similar work. We invite open-source software developers to join us in bringing VS-Hub to a much wider community.

Chapter 7. Conclusion

The primary findings in this research have been about the many ways value set development, sharing, and reuse can go wrong, and, at each step, recommendations are given for addressing these. Finally in Chapter 6, we present software implementing some of these recommendations and addressing reuse difficulties better than previous tools have done—a good beginning. The dissertation's findings, models, and conclusions expand greatly on prior understandings of the challenges besetting value set developers and the developers of methods and software to support them. The following pages briefly recap and summarize the trajectory of the research and synthesize the preceding chapters to provide a high-level view of the primary problems in value set management and approaches to solving them.

The single, central problem addressed in the research is that of finding or making value sets and then verifying that they are both appropriate and accurate for a given purpose and context, that is, that they are fit-for-use. The focus is then on platforms for value set authoring and reuse and the barriers to their effectiveness.

7.1 Problem definition

The dissertation's textual journey begins with the first published paper in Chapter 2 [18], which established the parameters and overall goals of my research moving forward. It contextualized the place of value sets in RWD research, explained the importance of reusing value sets, and problematized reuse in new ways. The barriers to reuse identified in that first paper were in the difficulty potential re-users encounter in attempting to find value sets to meet their clinical record selection needs, and, before that, in even believing that the benefit in finding a value set rather than defining a new one could justify the effort of finding one and verifying

that it is fit-for-use. Further, for researchers to contribute to repositories the value sets that they create or that they reuse and modify to make them most helpful to future re-users is a burdensome task, requiring more metadata and documentation than is commonly committed to value set sharing. To motivate that sharing will require it to be easier than to not share, and to make undeniably worthwhile use of the data shared.

At that point in the research, it appeared that the primary barriers to reuse were in finding and understanding reusable value sets and could be solved basically by making value set repositories more FAIR (see Section 4.4.5.) Value sets could be made more *findable*, *accessible*, and *interoperable* with harmonized standards, better metadata, centralized or federated repositories, and common repository software that could be integrated into disparate analytic platforms. Making them more *reusable* and verifiable could be done with new software features: the ability to compose value sets from Boolean operations on existing value sets; computably traceable provenance; capabilities to facilitate crowd-sourced curation[28] and network effects; semantic graph visualization; and linking from semantic graph to local patient data.

Chapter 2 assumes that users would be able to effectively explore and answer their own questions about fitness-for-use and differences between value sets with highly interactive, information-packed user interfaces, which would include features like visualization

- Of semantic relationships and hierarchies,
- For comparing similar value sets,
- For improving comprehensibility with parsimoniously structured intensional views on extensional or haphazardly intensional value sets,
- Of integrated code counts, value set counts, or other patient information.

[28] It only became clear later in the research that crowd-sourced curation actually makes value set repositories worse. It could only work, I believe, if a very high-quality repository became so popular that, as with Wikipedia, highly engaged volunteers were willing to monitor for and correct the mistakes of others.

Visualization and other user interface features for making as much relevant information as possible available to the user do go a long way toward getting users to make better decisions and leverage the work of prior value set developers—but these are not enough.

7.2 Fieldwork and conceptual models

In the field study presented in chapters 4 and 5, I learned a great deal about the specific practices of a diverse selection of value set developers and about their beliefs around the development, use, and reuse of value sets in RWD research projects. In the analysis, I grouped three dichotomies (permissible/analytic, prescriptive/descriptive, semantic/empirical) together as part of a conceptual model for understanding these findings—but I fear I have not yet made clear the justification for highlighting these particular distinctions and grouping them together. As to grouping them, there is a rough correspondence on the one hand between permissible value sets, prescriptive views of medical vocabulary use, and semantic-only approaches to value set authoring and validation, and on the other hand between analytic, descriptive, and empirical. But the import of the three dichotomies are very different from each other.

The import of the permissible/analytic distinction is for background understanding and as an aid to explaining the ecosystem around RWD research.[29] As a contribution to the literature, my hope is that future scholars will account for this distinction and specify whether their studies and findings apply mostly to data capture scenarios (permissible), RWD research (analytic), and/or other contexts.

[29] And also, to help readers understand the fundamental difference in the meaning of the two types. The user of a permissible value set is meant to choose one code out of a range of alternatives; there should be as little ambiguity between them as possible. With analytic, the codes are treated as equivalent. Records matched by any of them all go into the same bucket for analysis.

The prescriptive/descriptive distinction is sociologically interesting, but it is included to alert people to their own possible biases or blind spots and those of others. For those who work primarily with permissible value sets for data capture, it only makes sense that they would not consider using validation methods based on existing patient data. For those who work on RWD research, there is a continuum between those who rely exclusively on semantic methods and those who believe value sets or phenotypes *must* be validated by formal empirical methods based on the patient data to be analyzed. A belief that only semantic methods are necessary *might* be explained by a prescriptive bias, though it might also be explained by an awareness of obstacles to using patient data.

The distinction between semantic and empirical methods can seem strained; it applies not just to what people do but is defined by the data or resources they need to do it with. It was particularly helpful in organizing and presenting all the material gathered in the field study. Some comments from participants (P05 especially) led me to drawing this distinction, but the way I use the terms is idiosyncratic and has been confusing to some readers. I am not wedded to the terms semantic and empirical and would welcome future scholarship that could encapsulate these ideas with better terms.

That said, the distinction between methods that do or do not require patient-level data is essential for understanding the obstacles to value set validation and reuse, as well as for understanding all the various recommendations I make for advancing value set authoring and repository tools. Semantic methods and resources are necessary in developing or reusing any value set, permissible or analytic. Patient-level data and associated empirical methods are generally relevant only to RWD research and analytic value sets.

Since my work centers on RWD research, patient data are central, of course, and despite several pages of treatment (especially pages 92–105), I barely scratch the surface of the role of analysis and validation based on patient-level data in value set and phenotype development. But the reality seems to be that value sets are mostly authored and reused without empirical validation. Though the *best* authoring and repository software would support all the semantic and empirical methods, the best software that could feasibly be built for a wide audience in the near future would not attempt to integrate patient-level data. However, summary data (patient and record counts primarily) *should* be integrated as much as possible and should ideally accommodate counts from multiple sources.

Taken together, the three dichotomies point to the need for scholars in this space to situate themselves by declaring any philosophical leaning or bias they have, clarifying whether their findings apply to permissible or analytic value sets or both, and, for analytic value sets, reporting whether empirical data were available or used.

7.3 Vocabulary visualization and value set comparison

In recent years, ATLAS and the N3C concept set browser have added modest functionality for comparing one value set to similar ones. Neither make it easy to quickly or intuitively see the commonalities and differences amongst a group of similar value sets, and comparing metadata between value sets requires navigating between individual value set pages.

The central goal of VS-Hub (see Figure 6) is to *bring together vocabulary hierarchies and value set data into an integrated interface*. It demonstrates the value and technical feasibility of this goal to improve users' ability to leverage the knowledge embedded in existing value sets. In

a single view, it combines authoring features with rich information about the value sets chosen for comparison (definitions, expanded contents, and metadata) and their member codes. It allows similar value sets to be compared in groups rather than only against a single, focal value set. The indented concept table facilitates comparison at the level of individual codes and hierarchy subtrees, showing metadata and counts for each code and the value sets in which it was included, omitted, or explicitly excluded. As currently implemented, VS-Hub is still missing many features that would contribute toward its central goal. Nevertheless, it is a good start in showing what can be achieved with the data present in existing repositories.

The primary contribution of Chapter 6 is to show VS-Hub as a proof of concept and as an effort in research through design [160]. VS-Hub was developed in response to the active needs of the N3C community. The design and development team worked directly with (and served as) professionals responsible for the authoring or reuse, curation, and quality of value sets. In the process of designing and implementing VS-Hub, observing people (including ourselves) using it, and fielding questions and issues, much came to light about the practices and needs of value set authors and re-users; the features people want and use; mistakes commonly made in value set authoring; complexities in value set management that had not been on our radar; and what should be expected from state-of-the-art value set management platforms. Following are a few examples.

Working with large value sets is particularly difficult. On one occasion, a value set for malignant cancer required review by a clinical expert. The value set definition included six concepts: malignant neoplastic disease (which has 54,000 descendants) and five exclusions. In order to review the 54,000 concepts in this value set's expanded contents in the N3C concept set

browser or in ATLAS would be effectively impossible: the concepts might be sorted alphabetically or, in ATLAS, by record count, but not hierarchically, so the reviewer would need to page through hundreds of pages of concepts individually. Both tools allow contents to be downloaded as a spreadsheet, but, again, not helpfully ordered. In this case, actually, the reviewer did not have an N3C Enclave account and would have needed a spreadsheet export anyway.

At that point, with VS-Hub, it was not possible to browse around such a large value set without crashing the browser, but we did make it possible to download a spreadsheet with the concepts ordered and nested hierarchically. That allowed the reviewer to work on it, but the result of the review was a huge spreadsheet without a clear indication of what had been added and removed or any straightforward way to create a parsimonious new version by modifying the original six intensional definition rules. Instead, we would need to upload the new set of 54,300 concept codes, which, as it turned out, caused the N3C concept set editor to crash.

We learn several things from this story. People working on value sets do not necessarily have access to the repository where they are stored (VS-Hub is able to provide access without an account because it only contains semantic and count data.) The difficulties in working without the benefit of hierarchical nesting become insurmountable with large enough value sets—but with smaller value sets, even if no one complains to software developers, these difficulties may still be affecting the process and compromising review quality and results. Efforts to maintain or improve value sets may result in the replacement of intensional value sets with extensional ones (see footnote about the advantages of parsimonious value set definitions in Section 6.2.)

In our own team's work maintaining or reviewing value sets we had made, VS-Hub revealed mistakes we had not noticed when working with other tools. As a particular example, we saw many cases where a code would be marked for exclusion but without also being marked to include descendants. It is easy to be confused about this. Without include descendants, an excluded concept will be excluded, but its descendants will not be. Almost always, the intent would be to exclude descendants as well. Seeing this alerted us to the need for user alerts and explanatory text to prevent the mistake.

We learned that the reason other tools have not incorporated design innovations like those in VS-Hub is probably not because no one ever thought of them before but because there are very difficult to implement. As an educated guess, to build production-grade software implementing VS-Hub's current features along with others that should be included to meet VS-Hub's overall design goals might require between five and ten person years by a skilled software development team. I believe the benefits of such an investment would far outweigh the costs by dramatically increasing the efficiency and quality of value set development and reuse efforts.

It should be understood, though, that VS-Hub's design goals are a subset of the recommendations made in Chapter 2, before the field study. Although VS-Hub has taught us much about how to design value set management platforms, it has not incorporated what we learned from the fieldwork and analysis presented in Chapters 4 and 5. It constitutes a major advance in allowing users to explore *how* value sets differ from each other. In order for users to select the best value set for their purpose, they should ask, *Why is this value set different from all other value sets for this topic?*

7.3.1 Discerning reasons for value set differences

Redundancy and poor-quality value sets and the complexities in how same-topic value sets can differ from each other make it exceedingly difficult for users to find the best one for their purpose. Curatorial efforts far beyond anything tried at VSAC or N3C might succeed in paring a repository down to high-quality, nonredundant value sets, assuring that they have useful metadata, and maintaining quality as value sets are added—but it is hard for me to imagine any organization dedicating the resources this would require. Even if one could (magically) begin with only necessary variations on same-topic value sets, it would be difficult for curators to identify reasons for differences between them or to distinguish the kids of use each is fit for, much less to convey all this effectively to users.

The N3C concept set editor (see Section 6.1) collects metadata like intention, limitations, and provenance that are meant to help potential re-users understand whether a value set is fit for their use, but those data do not provide much guidance in distinguishing why value sets are different from each other. For one, value set creators often rush through the metadata prompts without a lot of thought. Even with lots of thought, they cannot be expected to know the information potential re-users really need to know, why this value set differs from all the similar ones. Even if that information were available, the interface as currently designed would require users to laboriously bring up each value set and remember its metadata in order to compare to the others.

There can be differences in clinical meaning or nuance; in study requirements like sensitivity and specificity, terminology or mapping issues, population or database characteristics, mistakes by value set authors, or judgment calls when keeping a code would result in false positives and omitting it would result in false negatives.

At a higher level, one could categorize differences as shown in Table 19. Similar, redundant, and junk value sets

Table 19. Similar, redundant, and junk value sets

SIMILAR / SAME TOPIC	REDUNDANT	JUNK
Same or similar name	Same or similar name	Created as an example or test by someone playing around or learning the system
Overlapping codes	Identical codes	Created as part of software testing
Justifiably different	Different codes due to error	Work-in-progress, never finished, or abandoned

SIMILAR / SAME TOPIC	REDUNDANT	JUNK
Same or similar name	Same or similar name	Created as an example or test by someone playing around or learning the system
Overlapping codes	Identical codes	Created as part of software testing
Justifiably different	Different codes due to error	Work-in-progress, never finished, or abandoned

Table 20. Why is a code in one value set and not another?Table 21. Value sets can be similar in name or contents but differ for legitimate reasons—in which case we would like to know what those reasons are so we can pick the best one for our purpose.

Or they can just be redundant: identical in contents, or similar in name or intent and having different codes, but not for any good reason, because codes were overlooked or included inappropriately.

And repositories can also accumulate junk value sets, created for testing or playing around with the software, or created and abandoned for some reason. Figuring out that a value set is junk by examining it is not easy, but the author probably knows. We might be able to tag junk value

sets to be removed or ignored by simply asking the author: Do you think this value set is a good candidate for reuse in future phenotypes or studies?

So, value sets on the same topic can be justifiably different, redundant, or junk. But if we want to find the most fit-for-use, we will need to go further in understanding reasons for differences.

To simplify, let us say a particular code is present in a value set made by A and absent in another made by B (Table 22. Why is a code in one value set and not another?

Value set A	Value set B
111	111
222	222
333	
444	444

Table 23. Difference discernment methodsTable 24.) We would like to know why. We would at least like to know, was A more thorough in collecting codes? Or was B aware of the code and intentionally omitted it? An expert with time on their hands might make an educated guess based

Table 22. Why is a code in one value set and not another?

Value set A	Value set B
111	111
222	222
333	
444	444

on a close examination of both value sets. But software could provide an answer if either 1) the code collection process of each had been captured and could be effectively presented to the potential re-user; or 2) B had given a reason for omitting the code.

Moving to the more detailed reasons classified in 5.3.6, I went through the exercise of trying to think how each type of difference might be discerned, whether computational reasoning could be applied or if it would require human interpretation or dedicated validation efforts. Table 25. Difference discernment methods

REASONS FOR CODE INCLUSION DIFFERENCES	DISCERNMENT METHOD	METADATA TO COLLECT
VALID SEMANTIC REASONS		
- Clinical meaning or nuance - Study requirements especially sensitivity and specificity - Algorithmic context	- Human interpretation of (new)metadata	- Specific questions comparing new value set to those already in the repository
- Terminologies and cross-terminology mappings - Use of vocabularies lacking granularity for clinical concepts or requiring post-coordination	- Human or computational use of (new) metadata - Alert re-user to known issues when mappings have been used	- Ask user when mappings with known issues have been used
VALID EMPIRICAL REASONS		
- Population characteristics - Database characteristics	- Empirical methods: comparison to gold standard, less systematic chart review, sensitivity analysis - Human or automated reasoning based on available validation data or specific knowledge - Evaluation steps of the CEER process	- Ask if empirical data were available, how used, for validation results
- Regional, institutional, or clinical specialization coding practices - Institutional workflow	- Human reasoning based on specific knowledge	- Optional question
ERRONEOUS OR ARBITRARY CAUSES		
- Codes mistakenly left out - Codes mistakenly included - Codes included or not based on faulty or idiosyncratic reference standards	- Empirical methods: comparison to gold standard, less systematic chart review, sensitivity analysis - Human or automated reasoning based on available validation data - Evaluation steps of the CEER process	- Prompt user to consider these possibilities, but they should fix rather than document them
- Arbitrary inclusion thresholds	- Consider impact if metadata has been captured	- Ask about codes that produce false positives when included, false negatives when excluded

Table 26 adds two additional columns to the reasons listed in Table 15.

Human interpretation might be done by the potential re-user, by the author of each value set, or by curators of the repositories. In any case, they would generally need more information at

their fingertips than they currently have access to. While interface and visualization improvements like those in VS-Hub should make it easier for expert users to guess at reasons for value set differences, I believe the only way for automated methods to effectively discern reasons or help the user do so would be with the capture of new metadata during authoring and validation processes.

Table 25. Difference discernment methods

REASONS FOR CODE INCLUSION DIFFERENCES	DISCERNMENT METHOD	METADATA TO COLLECT
VALID SEMANTIC REASONS		
- Clinical meaning or nuance - Study requirements especially sensitivity and specificity - Algorithmic context	- Human interpretation of (new)metadata	- Specific questions comparing new value set to those already in the repository
- Terminologies and cross-terminology mappings - Use of vocabularies lacking granularity for clinical concepts or requiring post-coordination	- Human or computational use of (new) metadata - Alert re-user to known issues when mappings have been used	- Ask user when mappings with known issues have been used
VALID EMPIRICAL REASONS		
- Population characteristics - Database characteristics	- Empirical methods: comparison to gold standard, less systematic chart review, sensitivity analysis - Human or automated reasoning based on available validation data or specific knowledge - Evaluation steps of the CEER process	- Ask if empirical data were available, how used, for validation results
- Regional, institutional, or clinical specialization coding practices - Institutional workflow	- Human reasoning based on specific knowledge	- Optional question
ERRONEOUS OR ARBITRARY CAUSES		
- Codes mistakenly left out - Codes mistakenly included - Codes included or not based on faulty or idiosyncratic reference standards	- Empirical methods: comparison to gold standard, less systematic chart review, sensitivity analysis - Human or automated reasoning based on available validation data - Evaluation steps of the CEER process	- Prompt user to consider these possibilities, but they should fix rather than document them
- Arbitrary inclusion thresholds	- Consider impact if metadata has been captured	- Ask about codes that produce false positives when included, false negatives when excluded

Collecting and making use of this additional metadata, however, would be a heavy lift for users. Software features would be designed to alleviate as much of the burden as possible, but, given the current state of repositories, with all the redundant and junk value sets they contain, asking the author why their value set is different from all similar ones is not going to be feasible. The number of comparisons that need to be made is on the order of N squared, which might be fine for computational analysis but not for humans attempting to understand and document all

these differences. That is, same-topic value sets could be kept to a minimum. Heroic efforts on the part of curators could be made to weed out junk and unnecessary redundancy, or one could start with an empty repository. So, if we could limit the burden on value set authors to only document differences from legitimately similar value sets, what kinds of metadata should be collected and how could they be used to help potential re-users understand the differences and choose the best for their needs?

Though study or phenotype requirements around sensitivity and specificity are mentioned throughout, previous chapters have lumped them together with other valid semantic reasons for difference. In practice, though, I have learned through observation and conversations mostly at N3C, sensitivity and specificity requirements are the most common valid reason that multiple value sets are needed for the same topic. Very specific value sets may be appropriate when recruiting for a study or selecting patients for a main cohort. Very sensitive value sets may be used to select patients for a screening test or when justifying possible impact to potential study funders. For covariates or studies of incidence or prevalence, a more balanced value set may be needed.

As discussed in Section 4.4.9.3, Refining versus validating code sets, calculating actual sensitivity and specificity values requires a gold standard of marked records of patients having and not having the condition of interest to compare value set/phenotype results against. As such standards are expensive to produce and seldom available, a lower bar would be systematic marking of records matching the value set as true or false positives. This would allow calculation of positive predictive value.

Most often, no empirical validation is available at all, so the question of sensitivity and specificity is one of intent rather than evidence. The N3C concept set editor already asks users whether their value sets are meant to be broad or narrow. It is not clear how thoughtfully that question tends to be answered or whether potential re-users pay attention to it. Automated methods might provide useful information on this count. If one value set contains a superset of the codes appearing in another, for instance, it must be more sensitive (at least in intent; whether in actuality depends on the additional codes appearing in patient records.)

Beyond sensitivity and specificity, there may be differences in *meaning*. For instance, a study might look at an allergic reaction like angioedema and need a value set for all the codes that indicate angioedema. But if the study was trying to determine the relative risks of drugs like ACE inhibitors, a very specific value set might only include codes for drug-induced angioedema. A more sensitive value set could include allergic angioedema, angioedema of lips, idiopathic or respiratory angioedema. But the value set should definitely exclude concepts like hereditary angioedema.

With such metadata, large language models (LLM) might help potential re-users by summarizing differences or by asking their intent and focusing their attention towards the most relevant value sets. Eventually, if a repository gathered enough explanatory text on meaningful differences amongst similar-topic value sets, an LLM might be trained to produce new explanatory text based solely on a value set's codes.

For *study requirements* and *algorithmic context*, perhaps phenotyping platforms could share information with value set analysis tools in helpful ways. But this information would, for the foreseeable future, need to be captured manually as unstructured text.

Where differences are due to *terminology* or *mapping* issues, visualization tools are already helpful, but could be more so if mapping and provenance history were captured computationally.

Differences due to *coding practices* or *clinical workflow* could possibly be explored empirically through large, multi-site studies whose results might be leveraged by value set analysis tools. Until then, terminology, informatics, and subject matter experts may have helpful ideas.

Differences due to *population or database characteristics* or to *errors or arbitrary causes* can be explored using the empirical techniques discussed in Section 4.4.9.

1. We have seen that Choosing the most fit-for-use value set from a number of alternatives is challenging, even without the presence of junk value sets;
2. The information available in current value set repositories is not enough to discern the reasons for differences;
3. Collecting additional metadata about these reasons during value set development could help; but
4. The design challenges in automating metadata collection and/or minimizing the burden on value set authors are formidable; and
5. May be intractable unless repositories can be kept free of junk and redundant value sets.

This *value-set-focused* approach to reuse (helping users understand differences between alternative value sets in order to identify the most fit-for-use) is not the only way to leverage the data and knowledge produced by prior value set authors, and I am more sanguine about the *code-focused* approach (presenting information about each code to best inform inclusion decisions) described in the next section. Yet, when people think of value set reuse, what they generally have in mind is that a specific value set will be reused, and value-set-focused reuse remains an important goal. Even with current repositories, curational efforts like the development and maintenance of an N3C-recommended library for commonly needed value sets can save time and improve quality as long as validation challenges and generalizability issues are recognized.

Further, the value-set-focused and code-focused approaches are not mutually exclusive and, eventually, I believe whole value sets will be accompanied by enough metadata to allow well-informed reuse. But the first steps on that path are large hurdles.

7.3.2 Code-focused reuse

Shifting reuse focus from value sets to codes is this dissertation's most radical contribution. Only a few participants spoke of systematically comparing sets of value sets and those who did built their own comparison routines in Python or SQL. This allowed them to focus on the codes, trying to discern why a given code would appear in some value sets and not others, bypassing questions of why the value sets are different from each other. (See Section 4.7.3.)

As with the value-set focused approach, the power to discern *why* a code or rule has been included or not will depend on new metadata collection but, even without this, VS-Hub already advances user understanding of *how* value sets differ by facilitating a more code-focused approach to value set comparison and authoring.

The indented list (Figure 6) shows each code's place in the hierarchy and most of the columns give information about the code. The value set columns on the right show each code's role in each of the selected value sets, which helps, of course, in understanding the codes as well as the differences and similarities between value sets.

But there is only enough room to select a handful of value sets. We might want to design an option for displaying summary data for each code based on all the value sets it appears in. For instance, the interface could show how many value sets include a given code explicitly, with or without descendants, or include it only as a descendant of some other included code, or exclude

it (exclusion is not the same as omission, by the way; it allows exclusion of codes that would otherwise be included as descendants of other codes.)

Or, it could be valuable to let the user explore a list of the codes this code most commonly co-occurs with. It might be particularly interesting to look at value sets that contain the greatest number of these commonly co-occurring codes but not the code itself.

There is much valuable work left to be done short of collecting more metadata, but understanding the *why*s rather than just the *how*s would be a game changer. With a code-focused approach, leveraging the work and knowledge of previous value set authors would mean knowing *why* a code has been included, excluded, or omitted. We could answer that question by asking authors for their reasoning as they are developing the value set. The CEER model is particularly designed to facilitate this by separating out code collection from code and value set evaluation. By making code collection a separate and prior activity, metadata can be gathered about why codes which seem worth consideration were omitted.

The interface design would need to minimize cognitive load or friction for the author by, for instance, prompting them with the most likely reasons. For inclusion, the default would probably be the perfunctory, "I'm including angioedema in this angioedema value set because that's what it's about." More interesting would be, for instance, "I'm including idiopathic angioedema because I suspect it is more often drug-induced than not," or "I'm omitting (or excluding) hereditary angioedema because it's not drug-induced." Future value set authors are likely to benefit especially from being alerted to intentional omission reasoning, in case it is relevant to their own purpose.

Code-level metadata might also be collected computationally: search strings, vocabulary navigation steps, or copying from another value set might be recorded when these actions lead directly to the inclusion of a code. It might also be possible, without requiring code collection and evaluation steps to be separated, to monitor search strings and navigation steps that do not lead to code collection and to record the codes viewed as having been omitted from that value set.

Such features would constitute a significant change in any interface that adopted them, but the data collected might prove invaluable. Unlike the value-set-focused approach, the presence of junk or redundant value sets in a repository would not be a hindrance to collecting metadata about value set authors' reasoning at the level of individual codes and inclusion rules. Collecting code-level metadata will require new software design, which is not trivial, but it would not require heroic curation efforts and the burden it would place on value set authors could be kept very manageable. (They might even welcome timely questions that help them clarify their thinking.)

As a final note, a piece of metadata that might go a long way toward alleviating repository quality issues would be to ask a value set developer when they finalize their contribution to the repository: Would you recommend this value set be reused by others? If so, then go on to ask about how this value set is different from all others. And, if not, exclude it from the value sets shown for possible reuse.

7.4 Contributions

In addition to the contributions mentioned in the earlier chapters (e.g., comprehensive taxonomy of value set development practices, conceptual and process models, visualization

methods), this last chapter has distilled and distinguished solution recommendations into value-set-focused and code-focused categories. The dissertation makes numerous original contributions in both categories.

Much prior work discusses value set reuse generally but does not cover redundancy and the need to compare similar value sets in terms of fitness-for-use. Though prior work does discuss the need for value set metadata, this, to my knowledge, is the first to discuss in detail the challenges involved in collecting various types of metadata and possibilities for overcoming these challenges.

Prior work does not mention the possibility of code-focused reuse, but there are two important efforts that offer techniques for code suggestions based on existing value sets: the PHOEBE code recommender system from Ostropolets, et al. [34,128,161] and "Mining Hierarchies and Similarity Clusters from Value Set Repositories" from Peterson, et al. [31]. These both aid value set developers in considering codes for inclusion and both of these methods would enhance value set authoring tools. The Peterson approach organizes the codes recommended into value set hierarchy clusters, which could be quite valuable, but, beyond a sunburst visualization, they do not discuss user interface methods for integrating the statistics their algorithm produces into a value set authoring system. Neither PHOEBE nor Peterson offer detailed code-focused data to inform code inclusion decisions.

The fieldwork presented here shows the potential of what can be learned by studying this branch of clinical research informatics, and much remains to be done. Survey results did report significant use of value set repositories and reuse of value sets made by others (Section 4.5), but further work is needed to explore what people mean by reuse, how they go about it, how it fits

into larger phenotype and study development processes, and how the reuse problems discussed in this dissertation affect it. It would be especially valuable to conduct observation studies of the authoring/reuse/validation process from start to finish. CEER and the other conceptual model elements (see especially Table 7) might guide observational protocols and alert researchers to important parts of the process.

We need a much better understanding of how empirical validation is or is not done in value set and phenotype development. When it does occur, does it follow normative procedures? I posit that much phenotype and value set development is done with informal empirical validation or none at all. Is this true? If so, what kinds of quality efforts are made?

The possibility has been suggested to me that differences in value set composition might actually not affect research study results very much. This does not seem likely to me but is obviously an important question. Individual researchers might answer it for themselves by performing a kind of sensitivity analysis: taking a number of alternative value sets for each of their phenotypes and running their study with each. Even if only testing alternatives for a single value set (and studies can use many), this could be tedious. Having software platforms to facilitate such sensitivity analyses would help. N3C experience has shown that value set variations can have very large effects on selection results, but the field would benefit from a published study of this kind of sensitivity analysis across a range of RWD studies, phenotypes, and value sets.

This dissertation's recommendations for collecting new metadata at the value set and code levels and providing highly interactive user interfaces for leveraging these metadata are original

contributions and point the way to a new generation of value set repository platforms and authoring tools.

Appendix A. Survey questions

1. Name
2. Email

   Please include an email address if you might be willing to participate in the next phase of this study, an interview over web conference.
3. Organization, institution, department
4. Title/Role
5. Specializations or degrees related to health care
6. Professional memberships or communities

   ☐ AMIA

   ☐ ISPE

   ☐ OHDSI

   ☐ HL7
7. If you are a researcher, what are the major topics of your research?
8. What sorts of health data do you work with regularly?

   ☐ Coded data with codes from standardized terminologies like ICD9, ICD10, SNOMED, MedDRA, NDC, RxNorm, LOINC, etc.

   ☐ Coded data with codes from local, institutional terminologies

   ☐ Clinical notes

   ☐ Radiology images

   ☐ Clinical trials data

   ☐ Genomic data

   ☐ Pathology data

   ☐ Administrative data

   ☐ Medical device data

   ☐ Other (please explain)

   F*ollow-up:* If your answer does not include coded data, please submit the survey now. (And thank you very much…)

For the remainder of the survey, please answer the questions specifically as they relate to your work around analyzing or supporting the analysis of coded patient data, that is, data in which clinical events and observations are stored as codes from controlled terminologies.

9. How many substantial studies or analysis projects might you work on in a year?

   ☐ 0 to 1

   ☐ 1 to 5

   ☐ Lots
10. Do you work with teams on these projects?

    ☐ Yes

    How many people are on the teams for typical projects? How many organizations participate in your analytic projects?

What roles do you play? (E.g., investigator, analyst, statistician, informaticist, clinical expert, epidemiologist, terminologist.)
What roles are played by others?

☐ No

11. What data sources and analytic software tools do you and your team use in performing health data analyses?

a.Data types

☐ Claims
☐ Electronic health records
☐ Clinical trials data
☐ Other. Please list particular products

b.Data sources

☐ Institutional clinical data warehouse
☐ VA
☐ CMS
☐ Vendor-supplied databases
☐ Synthetic or public datasets like SynPUF or MIMIC
☐ Other. Please list particular products

c.Data models

☐ OMOP
☐ PCORNet
☐ Sentinel
☐ i2b2
☐ claim forms
☐ local system
☐ Other. Please list particular products

d.Health data-focused analytic software tools

☐ OHDSI
☐ i2b2
☐ local system
☐ EPIC
☐ Cerner
☐ Oracle Clinical
☐ Other. Please list particular products

e.General purpose analytic software

☐ R
☐ SAS
☐ SQL database
☐ Tableau
☐ Other. Please list particular products

12. Do code lists play a part in your or your team's analytic workflow? That is, if you were, for instance, studying statin use in diabetic patients, would your analysis include, in some way, a list of diabetes codes and a list of statin codes?
    Follow-up: If not, please explain and submit survey

We will refer to this type of code list as a "clinical concept value set" from here on.

13. Are clinical concept value sets a distinct part of your workflow or an inseparable part of cohort definition and other analytic tasks?
    Follow-up: Please explain
14. Do you create clinical concept value sets, use ones created by others, or both?
15. Are the clinical concept value sets you use limited to a single vocabulary per domain (e.g., SNOMED or ICD10 for diagnoses, LOINC for labs, etc.) or do they draw from multiple vocabularies (e.g., SNOMED and ICD10 for diagnoses)?
16. How do you verify that you have selected the best codes for representing a clinical concept in your analyses?
17. What tools and resources do you make regular use of in developing or working with clinical concept value sets?
18. Do you ever develop clinical concept value sets that you expect or hope will be used by others?
    Follow-up: What do you consider best practices to facilitate discovery and reuse of your value sets by others?
19. What would you change in the tools you use for clinical concept value set discovery, authoring, editing, and evaluation?
20. Do you see a need for new tools for value set and terminology visualization?
21. How would such tools need to fit into your data analytics workflow?

Appendix B. Questions for code set development

0. Should I look at other code sets before constructing my own from scratch?
    1. Where should I look for existing code sets?
    2. What search criteria should I use?
    3. What do I do if I find many different code sets that seem to match my criteria?
        - Do I have a way of knowing which ones fit my use case better than others?
        - Can I distinguish one of them to reuse?
        - Should I combine them all into a single list?
    4. If I do manage to find and decide on a single or combined code set that matches my criteria:
        - Is it in a data format I can use?
        - Does it use the terminologies I care about?
        - Is it based on versions of the source vocabularies appropriate to my data?
        - Does it use those terminologies consistently with the way my target data use them?
        - Is this code set credible?
        - Is the repository it comes from recognized and accepted in my scholarly/professional community?
        - Am I trying to actually reuse this code set, or do I simply use the codes as a starting point?

    - If starting point, was it worth the work I've done to this point or would it have been easier to find these codes in the vocabularies myself?
- Has any evidence of how the code set was constructed and validated been made available to me?
    - Was consideration and selection of codes influenced in any way by phenotype (or study) characteristics? Or was the code set treated as a self-sufficient algorithm for matching appropriate patient records?
    - If code set performance was checked by review (formal or not) of database results, were those results generated by a containing phenotype algorithm? (Empirical)
    - If code sets are semantically or empirically "validated" separately from the phenotypes they will be contained in, why? Simply because that's how the process is set up? Because code set authoring and algorithm authoring are different skill sets not usually combined in the same person? Might this be problematic?
    - What database(s) was the code set created for?
    - How did they make inclusion/exclusion decisions?
    - What database(s) was it evaluated with? How?
    - Did they validate against a (gold or silver) reference standard?
        - Do I know how they constructed their reference standard?
            - MRA? How did they ensure accuracy? What training did the abstractors have [135,162]
            - With a random sample of the general database population? A random sample of a subset? How was the subset chosen?
    - Did they use random chart review?
    - What semantic evaluation/validation methods were used?
- Do I need to do any less validation than I would if building my own code set from scratch?
- Are there reasons I could trust prior evaluation or validation work, or do I need to do my own from scratch?

Bibliography

[1] J.P. Jarow, L. LaVange, J. Woodcock, Multidimensional Evidence Generation and FDA Regulatory Decision Making: Defining and Using "Real-World" Data, JAMA 318 (2017) 703. https://doi.org/10.1001/jama.2017.9991.

[2] 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.

[3] R. Platt, J.S. Brown, M. Robb, M. McClellan, R. Ball, M.D. Nguyen, R.E. Sherman, The FDA Sentinel Initiative - An Evolving National Resource, N Engl J Med 379 (2018) 2091–2093. https://doi.org/10.1056/NEJMp1809643.

[4] C. for D. and R. Health, Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices, US Food Drug Adm. (2019). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices (accessed December 7, 2020).

[5] 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 Med. 12 (2015) e1001885. https://doi.org/10.1371/journal.pmed.1001885.

[6] S. Schneeweiss, J.S. Brown, A. Bate, G. Trifirò, D.B. Bartels, Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products, Clin. Pharmacol. Ther. 107 (2020) 827–833. https://doi.org/10.1002/cpt.1577.

[7] OHDSI, The Book of OHDSI, 2020th-04–16th ed., Observational Health Data Sciences and Informatics, 2020. http://book.ohdsi.org.

[8] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers, Stud Health Technol Inf. 216 (2015) 574–578. https://www.ncbi.nlm.nih.gov/pubmed/26262116.

[9] Observational Health Data Sciences and Informatics, Publication analysis: PubMed OHDSI manuscripts, OHDSI Community Dashboard (n.d.). https://dash.ohdsi.org/pubmed (accessed June 20, 2023).

[10] 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, J. Am. Med. Inform. Assoc. 28 (2021) 427–443. https://doi.org/10.1093/jamia/ocaa196.

[11] National COVID-19 Cohort Collaborative, N3C Publications, N3C Dashboards (n.d.). https://covid.cd2h.org/dashboard/publications (accessed June 20, 2023).

[12] J.C. Maro, R. Platt, J.H. Holmes, B.L. Strom, S. Hennessy, R. Lazarus, J.S. Brown, Design of a National Distributed Health Data Network, Ann. Intern. Med. 151 (2009) 341–344. https://doi.org/10.7326/0003-4819-151-5-200909010-00139.

[13] L.H. Curtis, M.G. Weiner, D.M. Boudreau, W.O. Cooper, G.W. Daniel, V.P. Nair, M.A. Raebel, N.U. Beaulieu, R. Rosofsky, T.S. Woodworth, J.S. Brown, Design considerations, architecture, and use of the Mini-Sentinel distributed data system, Pharmacoepidemiol Drug Saf 21 Suppl 1 (2012) 23–31. https://doi.org/10.1002/pds.2336.

[14] R.L. Fleurence, L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby, J.S. Brown, Launching PCORnet, a national patient-centered clinical research network, J. Am. Med. Inform. Assoc. 21 (2014) 578–582. https://doi.org/10.1136/amiajnl-2014-002747.

[15] 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 Inf. Assoc 20 (2013) e319-26. https://doi.org/10.1136/amiajnl-2013-001952.

[16] 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.

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

[18] 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).

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

[20] 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, (2021) 35. https://doi.org/10.1101/2021.10.14.21264917.

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

[22] G. Hripcsak, D.J. Albers, High-fidelity phenotyping: richness and freedom from bias, J. Am. Med. Inform. Assoc. 25 (2018) 289–294. https://doi.org/10.1093/jamia/ocx110.

[23] 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, J. Am. Med. Inform. Assoc. 27 (2020) 1968–1976. https://doi.org/10.1093/jamia/ocaa200.

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

[25] 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).

[26] 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, Learn. Health Syst. 6 (2022) e10271. https://doi.org/10.1002/lrh2.10271.

[27] 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.

[28] 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).

[29] 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.

[30] 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.

[31] 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.

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

[33] 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, J. Biomed. Inform. 100 (2019) 103322. https://doi.org/10.1016/j.jbi.2019.103322.

[34] 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.).

[35] 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.

[36] 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.

[37] 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.

[38] 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).

[39] 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).

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

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

[42] 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. AMIA Symp. 2019 (2019) 438–447.

[43] 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.

[44] 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.

[45] The All of Us Research Program Investigators, The "All of Us" Research Program, N. Engl. J. Med. 381 (2019) 668–676. https://doi.org/10.1056/NEJMsr1809937.

[46] J.J. Cimino, Desiderata for controlled medical vocabularies in the twenty-first century, Methods Inf Med 37 (1998) 394–403.

[47] 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.

[48] P.E. Stang, P.B. Ryan, J.A. Racoosin, J.M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, J. Woodcock, Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership, Ann Intern Med 153 (2010) 600–606. https://doi.org/10.7326/0003-4819-153-9-201011020-00010.

[49] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, P.E. Stang, Validation of a common data model for active safety surveillance research, J Am Med Inf. Assoc 19 (2012) 54–60. https://doi.org/10.1136/amiajnl-2011-000376.

[50] R.L. Fleurence, L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby, J.S. Brown, Launching PCORnet, a national patient-centered clinical research network, J. Am. Med. Inform. Assoc. 21 (2014) 578–582. https://doi.org/10.1136/amiajnl-2014-002747.

[51] 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).

[52] S. Gold, H.P. Lehmann, L.M. Schilling, W.G. Lutters, Value sets and the problem of redundancy in value set repositories, 2024. https://doi.org/10.1101/2024.02.15.24302903.

[53] G. Jiang, H.R. Solbrig, C.G. Chute, Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network, J Biomed Inf. 44 Suppl 1 (2011) S78-85. https://doi.org/10.1016/j.jbi.2011.08.001.

[54] G. Jiang, R.C. Kiefer, D.K. Sharma, E. Prud'hommeaux, H.R. Solbrig, A Consensus-Based Approach for Harmonizing the OHDSI Common Data Model with HL7 FHIR, Stud Health Technol Inf. 245 (2017) 887–891. https://www.ncbi.nlm.nih.gov/pubmed/29295227.

[55] K.J. Peterson, G. Jiang, S.M. Brue, H. Liu, Leveraging Terminology Services for Extract-Transform-Load Processes: A User-Centered Approach, AMIA Annu Symp Proc 2016 (2016) 1010–1019. https://www.ncbi.nlm.nih.gov/pubmed/28269898.

[56] R.L. Richesson, J. Sun, J. Pathak, A.N. Kho, J.C. Denny, Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods, Artif Intell Med 71 (2016) 57–61. https://doi.org/10.1016/j.artmed.2016.05.005.

[57] H. Mo, G. Jiang, J.A. Pacheco, R. Kiefer, L.V. Rasmussen, J. Pathak, J.C. Denny, W.K. Thompson, A Decompositional Approach to Executing Quality Data Model Algorithms on the i2b2 Platform, AMIA Jt Summits Transl Sci Proc 2016 (2016) 167–175. https://www.ncbi.nlm.nih.gov/pubmed/27570665.

[58] J.S. Brown, M. Kahn, S. Toh, Data quality assessment for comparative effectiveness research in distributed data networks, Med Care 51 (2013) S22-9. https://doi.org/10.1097/MLR.0b013e31829b1e2c.

[59] S.T. Rosenbloom, R.J. Carroll, J.L. Warner, M.E. Matheny, J.C. Denny, Representing Knowledge Consistently Across Health Systems, Yearb Med Inf. 26 (2017) 139–147. https://doi.org/10.15265/IY-2017-018.

[60] H. Mo, W.K. Thompson, L.V. Rasmussen, J.A. Pacheco, G. Jiang, R. Kiefer, Q. Zhu, J. Xu, E. Montague, D.S. Carrell, T. Lingren, F.D. Mentch, Y. Ni, F.H. Wehbe, P.L. Peissig, G. Tromp, E.B. Larson, C.G. Chute, J. Pathak, J.C. Denny, P. Speltz, A.N. Kho, G.P. Jarvik, C.A. Bejan, M.S. Williams, K. Borthwick, T.E. Kitchner, D.M. Roden, P.A. Harris, Desiderata for computable representations of electronic health records-driven phenotype algorithms, J Am Med Inf. Assoc 22 (2015) 1220–1230. https://doi.org/10.1093/jamia/ocv112.

[61] J.C. Kirby, P. Speltz, L.V. Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, S.B. Ellis, T. Lingren, W.K. Thompson, G.

Savova, J. Haines, D.M. Roden, P.A. Harris, J.C. Denny, PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inform. Assoc. JAMIA 23 (2016) 1046–1052. https://doi.org/10.1093/jamia/ocv202.

[62] Centers for Medicare & Medicaid Services, Office of the National Coordinator for Health Information Technology, Statin Therapy for the Prevention and Treatment of Cardiovascular Disease, (2017). https://ecqi.healthit.gov/ecqm/measures/cms347v1 (accessed March 3, 2018).

[63] F.R. Goss, L. Zhou, J.M. Plasek, C. Broverman, G. Robinson, B. Middleton, R.A. Rocha, Evaluating standard terminologies for encoding allergy information, J Am Med Inf. Assoc 20 (2013) 969–979. https://doi.org/10.1136/amiajnl-2012-000816.

[64] R.L. Richesson, P. Nadkarni, Data standards for clinical research data collection forms: current status and challenges, J Am Med Inf. Assoc 18 (2011) 341–346. https://doi.org/10.1136/amiajnl-2011-000107.

[65] A.L. Rector, R. Qamar, T. Marley, Binding ontologies and coding systems to electronic health records and messages, Appl Ontol 4 (2009) 51–69.

[66] S.W. Tu, J.R. Campbell, J. Glasgow, M.A. Nyman, R. McClure, J. McClay, C. Parker, K.M. Hrabak, D. Berg, T. Weida, Others, The SAGE Guideline Model: achievements and overview, J Am Med Inf. Assoc 14 (2007) 589–598. https://academic.oup.com/jamia/article-abstract/14/5/589/720934.

[67] A.L. Rector, What's in a code? Towards a formal account of the relation of ontologies and coding systems, Stud Health Technol Inf. 129 (2007) 730–734. https://www.ncbi.nlm.nih.gov/pubmed/17911813.

[68] S.J. Reisinger, P.B. Ryan, D.J. O'Hara, G.E. Powell, J.L. Painter, E.N. Pattishall, J.A. Morris, Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases, J Am Med Inf. Assoc 17 (2010) 652–662. https://doi.org/10.1136/jamia.2009.002477.

[69] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, M.N. Zozus, Evaluating common data models for use with a longitudinal community registry, J Biomed Inf. 64 (2016) 333–341. https://doi.org/10.1016/j.jbi.2016.10.016.

[70] M.G. Kahn, 04-EHR data methodologies in clinical research: perspectives from the field. Session 1: Semantic harmonization: definition; content; ontologies. Common data models for sharing EHR data across settings, Health Sci. Libr. Photogr. Collect. Spec. Collect. Univ. Colo. Anschutz Med. Campus Health Sci. Libr. Ser. V Sch. Med. Publ. (2007). https://dspace.library.colostate.edu/handle/10968/737.

[71] B.M. Kuehn, FDA's Foray Into Big Data Still Maturing, JAMA 315 (2016) 1934–1936. https://doi.org/10.1001/jama.2016.2752.

[72] P. Velentgas, R.L. Bohn, J.S. Brown, K.A. Chan, P. Gladowski, C.N. Holick, J.M. Kramer, C. Nakasato, C.M. Spettell, A.M. Walker, F. Zhang, R. Platt, A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study, Pharmacoepidemiol. Drug Saf. 17 (2008) 1226–1234. https://doi.org/10.1002/pds.1675.

[73] J.S. Brown, M. Kulldorff, K.A. Chan, R.L. Davis, D. Graham, P.T. Pettus, S.E. Andrade, M.A. Raebel, L. Herrinton, D. Roblin, D. Boudreau, D. Smith, J.H. Gurwitz, M.J. Gunter,

R. Platt, Early detection of adverse drug events within population-based health networks: application of sequential testing methods, Pharmacoepidemiol. Drug Saf. 16 (2007) 1275–1284. https://doi.org/10.1002/pds.1509.

[74] S. Schneeweiss, A basic study design for expedited safety signal evaluation based on electronic healthcare data, Pharmacoepidemiol. Drug Saf. 19 (2010) 858–868. https://doi.org/10.1002/pds.1926.

[75] V. Huser, J.J. Cimino, Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories, AMIA Annu Symp Proc 2013 (2013) 648–656. https://www.ncbi.nlm.nih.gov/pubmed/24551366.

[76] S. Bakken, An informatics infrastructure is essential for evidence-based practice, J Am Med Inf. Assoc 8 (2001) 199–201. https://www.ncbi.nlm.nih.gov/pubmed/11320064.

[77] G.C. Bowker, S.L. Star, Building information infrastructures for social worlds—The role of classifications and standards, in: Community Comput. Support Syst., Springer, 1998: pp. 231–248. https://link.springer.com/chapter/10.1007/3-540-49247-X_16.

[78] L.B. Becnel, S. Hastak, W. Ver Hoef, R.P. Milius, M. Slack, D. Wold, M.L. Glickman, B. Brodsky, C. Jaffe, R. Kush, E. Helton, BRIDG: a domain information model for translational and clinical protocol-driven research, J Am Med Inf. Assoc 24 (2017) 882–890. https://doi.org/10.1093/jamia/ocx004.

[79] F.J. DeFalco, P.B. Ryan, M. Soledad Cepeda, Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure, Health Serv Outcomes Res Methodol 13 (2013) 58–67. https://doi.org/10.1007/s10742-012-0102-1.

[80] L.H. Curtis, M.G. Weiner, D.M. Boudreau, W.O. Cooper, G.W. Daniel, V.P. Nair, M.A. Raebel, N.U. Beaulieu, R. Rosofsky, T.S. Woodworth, J.S. Brown, Design considerations, architecture, and use of the Mini-Sentinel distributed data system, Pharmacoepidemiol Drug Saf 21 Suppl 1 (2012) 23–31. https://doi.org/10.1002/pds.2336.

[81] F. DeFalco, OHDSI Architecture Workgroup, (n.d.). http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:architecture_wg (accessed March 3, 2018).

[82] J.J. Cimino, In defense of the Desiderata, J. Biomed. Inform. 39 (2006) 299–306. https://doi.org/10.1016/j.jbi.2005.11.008.

[83] 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).

[84] J. Lanier, Who Owns the Future?, Simon and Schuster, 2014. https://market.android.com/details?id=book-obDsAgAAQBAJ.

[85] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, P.E. Stang, Validation of a common data model for active safety surveillance research, J Am Med Inf. Assoc 19 (2012) 54–60. https://doi.org/10.1136/amiajnl-2011-000376.

[86] Observational Health Data Sciences and Informatics, The Book of OHDSI, 2020th-04–16th ed., 2020. http://book.ohdsi.org (accessed June 17, 2020).

[87] L. Ennis, T. Wykes, Impact of patient involvement in mental health research: longitudinal study, Br. J. Psychiatry 203 (2013) 381–386. https://doi.org/10.1192/bjp.bp.112.119818.

[88] S. Visweswaran, M.J. Becich, V.S. D'Itri, E.R. Sendro, D. MacFadden, N.R. Anderson, K.A. Allen, D. Ranganathan, S.N. Murphy, E.H. Morrato, H.A. Pincus, R. Toto, G.S. Firestein, L.M. Nadler, S.E. Reis, Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network, JAMIA Open 1 (2018) 147–152. https://doi.org/10.1093/jamiaopen/ooy033.

[89] the eMERGE Team, C.A. McCarty, R.L. Chisholm, C.G. Chute, I.J. Kullo, G.P. Jarvik, E.B. Larson, R. Li, D.R. Masys, M.D. Ritchie, D.M. Roden, J.P. Struewing, W.A. Wolf, The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies, BMC Med. Genomics 4 (2011). https://doi.org/10.1186/1755-8794-4-13.

[90] 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, J. Am. Med. Inform. Assoc. 18 (2011) 376–386. https://doi.org/10.1136/amiajnl-2010-000061.

[91] 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 Med. 12 (2015) e1001885. https://doi.org/10.1371/journal.pmed.1001885.

[92] 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.

[93] Rachel Richesson, PhD, MPH, Michelle Smerek, Shelley Rusincovitch, Meredith Nahm Zozus, PhD, Paramita Saha Chaudhuri, PhD, W. Ed Hammond, PhD, Robert M. Califf, MD, Greg Simon, MD, Beverly Green, MD, MPH, Michael Kahn, MD, PhD, Reesa Laws, BS, Electronic Health Records-Based Phenotyping, in: Rethink. Clin. Trials Living Textb. Pragmatic Clin. Trials, NIH Health Care Systems Research Collaboratory, Bethesda, MD, 2014. https://rethinkingclinicaltrials.org/resources/ehr-phenotyping/.

[94] 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. AMIA Symp. (2008) 556–560.

[95] R. Williams, E. Kontopantelis, I. Buchan, N. Peek, Clinical code set engineering for reusing EHR data for research: A review, J. Biomed. Inform. 70 (2017) 1–13.

[96] R.L. Richesson, W.E. Hammond, M. Nahm, D. Wixted, G.E. Simon, J.G. Robinson, A.E. Bauck, D. Cifelli, M.M. Smerek, J. Dickerson, Others, Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory, J Am Med Inf. Assoc 20 (2013) e226–e231. https://academic.oup.com/jamia/article-abstract/20/e2/e226/2909215.

[97] OHDSI/Atlas, Observational Health Data Sciences and Informatics, 2020. https://github.com/OHDSI/Atlas (accessed May 5, 2020).

[98] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, AMIA Annu Symp Proc (2006) 1040. https://www.ncbi.nlm.nih.gov/pubmed/17238659.

[99] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J Am Med Inf. Assoc 17 (2010) 124–130. https://doi.org/10.1136/jamia.2009.000893.

[100] HL7 Standard: Clinical Quality Language Specification, (n.d.). https://cql.hl7.org/ (accessed July 3, 2020).

[101] J.C. Kirby, P. Speltz, L.V. Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, S.B. Ellis, T. Lingren, W.K. Thompson, G. Savova, J. Haines, D.M. Roden, P.A. Harris, J.C. Denny, PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inform. Assoc. JAMIA 23 (2016) 1046–1052. https://doi.org/10.1093/jamia/ocv202.

[102] 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. JAMIA 19 (2012) 225–234. https://doi.org/10.1136/amiajnl-2011-000456.

[103] L.G. Hemkens, E.I. Benchimol, S.M. Langan, M. Briel, B. Kasenda, J.-M. Januel, E. Herrett, E. von Elm, The reporting of studies using routinely collected health data was often insufficient, J. Clin. Epidemiol. 79 (2016) 104–111. https://doi.org/10.1016/j.jclinepi.2016.06.005.

[104] 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, Learn. Health Syst. n/a (n.d.) e10271. https://doi.org/10.1002/lrh2.10271.

[105] 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. AMIA Symp. (2008) 556–560.

[106] 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).

[107] J.P. Vandenbroucke, E. von Elm, D.G. Altman, P.C. Gøtzsche, C.D. Mulrow, S.J. Pocock, C. Poole, J.J. Schlesselman, M. Egger, STROBE Initiative, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration, PLoS Med. 4 (2007) e297. https://doi.org/10.1371/journal.pmed.0040297.

[108] M. Conway, R.L. Berg, D. Carrell, J.C. Denny, A.N. Kho, I.J. Kullo, J.G. Linneman, J.A. Pacheco, P. Peissig, L. Rasmussen, N. Weston, C.G. Chute, J. Pathak, Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms, AMIA Annu. Symp. Proc. AMIA Symp. 2011 (2011) 274–283.

[109] L.K. Wiley, J.D. Moretz, J.C. Denny, J.F. Peterson, W.S. Bush, Phenotyping Adverse Drug Reactions: Statin-Related Myotoxicity, AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci. 2015 (2015) 466–470.

[110] S.C. Ahalt, C.G. Chute, K. Fecho, G. Glusman, J. Hadlock, C.O. Taylor, E.R. Pfaff, P.N. Robinson, H. Solbrig, C. Ta, N. Tatonetti, C. Weng, Clinical Data: Sources and Types, Regulatory Constraints, Applications, Clin. Transl. Sci. 12 (2019) 329–333. https://doi.org/10.1111/cts.12638.

[111] 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).

[112] A. Elixhauser, C. Steiner, Most common diagnoses and procedures in US community hospitals, 1996, Healthc. Cost Util. Proj. HCUP Res. Note Rockv. MD Agency Health Care Policy Res. AHCPR Pub (1999) 99–0046. https://www.hcup-us.ahrq.gov/reports/natstats/commdx/commdx.htm (accessed April 13, 2022).

[113] K. Cross, A. Warmack, Contextual inquiry: Quantification and use in videotaped analysis, in: Ext. Abstr. ACM Conf. Hum. Factors Comput. Syst., ACM, New York, NY, USA, 2000: pp. 317–318.

[114] D. Wixon, K. Holtzblatt, S. Knox, Contextual design: An emergent view of system design, in: Proc. ACM Conf. Hum. Factors Comput. Syst., ACM, New York, NY, USA, 1990: pp. 329–336. https://doi.org/10.1145/97243.97304.

[115] J. Lazar, Research methods in human computer interaction, 2nd edition, Elsevier, Cambridge, MA, 2017.

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

[117] E.I. Benchimol, D.G. Manuel, T. To, A.M. Griffiths, L. Rabeneck, A. Guttmann, Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data, J. Clin. Epidemiol. 64 (2011) 821–829. https://doi.org/10.1016/j.jclinepi.2010.10.006.

[118] S.V. Wang, S. Schneeweiss, M.L. Berger, J. Brown, F. de Vries, I. Douglas, J.J. Gagne, R. Gini, O. Klungel, C.D. Mullins, M.D. Nguyen, J.A. Rassen, L. Smeeth, M. Sturkenboom, Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0, Value Health 20 (2017) 1009–1022. https://doi.org/10.1016/j.jval.2017.08.3018.

[119] S.M. Langan, S.A. Schmidt, K. Wing, V. Ehrenstein, S.G. Nicholls, K.B. Filion, O. Klungel, I. Petersen, H.T. Sorensen, W.G. Dixon, A. Guttmann, K. Harron, L.G. Hemkens, D. Moher, S. Schneeweiss, L. Smeeth, M. Sturkenboom, E. von Elm, S.V. Wang, E.I. Benchimol, The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE), BMJ (2018) k3532. https://doi.org/10.1136/bmj.k3532.

[120] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J.

Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016). https://doi.org/10.1038/sdata.2016.18.

[121] A. Bosca, M. Casu, M. Dragoni, A. Rexha, Modeling, managing, exposing, and linking ontologies with a wiki-based tool, in: Proc. LREC, 2014: p. 1668.

[122] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267D – 270. https://doi.org/10.1093/nar/gkh061.

[123] C. Office of the National Coordinator for Health Information Technology, Common Data Model Harmonization: Harmonization of Various Common Data Models and Open Standards for Evidence Generation to Support Patient-Centered Outcomes Research, 2020. https://www.healthit.gov/sites/default/files/page/2020-07/CDMH-Project-Summary.pdf (accessed September 16, 2021).

[124] CDMH Team Members, Common Data Model Harmonization (CDMH) and Open Standards for Evidence Generation Final Report, FDA, NIH, ONC, 2020. https://aspe.hhs.gov/sites/default/files/private/pdf/259016/CDMH-Final-Report-14August2020.pdf (accessed September 16, 2021).

[125] HL7 Standards Product Brief - HL7 Version 3 Standard: Common Terminology Services (CTS), Release 2 | HL7 International, (n.d.). https://www.hl7.org/implement/standards/product_brief.cfm?product_id=384 (accessed February 24, 2023).

[126] C.G. Chute, P.L. Elkin, D.D. Sherertz, M.S. Tuttle, Desiderata for a clinical terminology server, Proc AMIA Symp (1999) 42–46. https://www.ncbi.nlm.nih.gov/pubmed/10566317.

[127] K.A. Shefchek, N.L. Harris, M. Gargano, N. Matentzoglu, D. Unni, M. Brush, D. Keith, T. Conlin, N. Vasilevsky, X.A. Zhang, J.P. Balhoff, L. Babb, S.M. Bello, H. Blau, Y. Bradford, S. Carbon, L. Carmody, L.E. Chan, V. Cipriani, A. Cuzick, M. Della Rocca, N. Dunn, S. Essaid, P. Fey, C. Grove, J.-P. Gourdine, A. Hamosh, M. Harris, I. Helbig, M. Hoatlin, M. Joachimiak, S. Jupp, K.B. Lett, S.E. Lewis, C. McNamara, Z.M. Pendlington, C. Pilgrim, T. Putman, V. Ravanmehr, J. Reese, E. Riggs, S. Robb, P. Roncaglia, J. Seager, E. Segerdell, M. Similuk, A.L. Storm, C. Thaxon, A. Thessen, J.O.B. Jacobsen, J.A. McMurry, T. Groza, S. Köhler, D. Smedley, P.N. Robinson, C.J. Mungall, M.A. Haendel, M.C. Munoz-Torres, D. Osumi-Sutherland, The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species, Nucleic Acids Res. 48 (2020) D704–D715. https://doi.org/10.1093/nar/gkz997.

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

[129] N. Vasilevsky, S. Essaid, N. Matentzoglu, N.L. Harris, M. Haendel, P. Robinson, C.J. Mungall, Mondo Disease Ontology: harmonizing disease concepts across the world, in: CEUR-WS, 2020.

[130] 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.

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

[132] 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, J. Am. Med. Inform. Assoc. (2021) ocab217. https://doi.org/10.1093/jamia/ocab217.

[133] I.A.O.C. Odysseus Data Services, Athena-OHDSI Vocabularies Hierarchy with Aggregation, (n.d.). http://athena.ohdsi.org/search-terms/terms/4101796/graph?levels=10&standardsOnly=false&zoomLevel=3 (accessed May 14, 2018).

[134] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of clinical and genetic data in the i2b2 architecture, AMIA Annu Symp Proc (2006) 1040. https://www.ncbi.nlm.nih.gov/pubmed/17238659.

[135] M.N. Zozus, C. Pieper, C.M. Johnson, T.R. Johnson, A. Franklin, J. Smith, J. Zhang, Factors Affecting Accuracy of Data Abstracted from Medical Records, PLOS ONE 10 (2015) e0138649. https://doi.org/10.1371/journal.pone.0138649.

[136] S. Gold, T. Zhang, R.L. Zhu, S. Hong, H.P. Lehmann, D. Gabriel, T. Francis, L. Eskenazi, C.G. Chute, ICD10–SNOMED mapping pitfalls: Post-coordinated expressions and concept sets, in: 2022 OHDSI Symp. Collab. Showc., Bethesda, Maryland, 2022. https://www.ohdsi.org/2022showcase-21/.

[137] G. Hripcsak, P.B. Ryan, J.D. Duke, N.H. Shah, R.W. Park, V. Huser, M.A. Suchard, M.J. Schuemie, F.J. DeFalco, A. Perotte, J.M. Banda, C.G. Reich, L.M. Schilling, M.E. Matheny, D. Meeker, N. Pratt, D. Madigan, Characterizing treatment pathways at scale

using the OHDSI network, Proc. Natl. Acad. Sci. 113 (2016) 7329–7336. https://doi.org/10.1073/pnas.1510502113.

[138] J.W. Tukey, Exploratory data analysis, Reading, Mass., 1977. http://theta.edu.pl/wp-content/uploads/2012/10/exploratorydataanalysis_tukey.pdf.

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

[140] 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 Inf. Assoc 19 (2012) e129-36. https://doi.org/10.1136/amiajnl-2011-000739.

[141] 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.

[142] 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).

[143] 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.), Encycl. Database Syst., Springer US, Boston, MA, 2009: pp. 1724–1727. https://doi.org/10.1007/978-0-387-39940-9_907.

[144] National Cancer Institute, caDSR and ISO 11179, Cancer Data Stand. Regist. Repos. (2016). https://wiki.nci.nih.gov/display/caDSR/caDSR+and+ISO+11179 (accessed May 26, 2020).

[145] 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.

[146] 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. JAMIA 20 (2013) e147–e154. https://doi.org/10.1136/amiajnl-2012-000896.

[147] 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 Med. Inform. Decis. Mak. 19 (2019) 31. https://doi.org/10.1186/s12911-019-0750-y.

[148] M.L. Markus, T. Connolly, Why CSCW applications fail: Problems in the adoption of interdependent work tools, in: Proc. 1990 ACM Conf. Comput.-Support. Coop. Work, 1990: pp. 371–380.

[149] M.L. Markus, Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success, J. Manag. Inf. Syst. 18 (2001) 57–93. https://doi.org/10.1080/07421222.2001.11045671.

[150] A. Perer, B. Shneiderman, Systematic yet flexible discovery: guiding domain experts through exploratory data analysis, in: Proc. 13th Int. Conf. Intell. User Interfaces, Association for Computing Machinery, New York, NY, USA, 2008: pp. 109–118. https://doi.org/10.1145/1378773.1378788.

[151] 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, Pharmacoepidemiol. Drug Saf. 18 (2009) 704–707. https://doi.org/10.1002/pds.1770.

[152] 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.

[153] 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).

[154] S. Gold, Value sets and the problem of redundancy in value set repositories. Survey data, (2024). https://doi.org/10.17605/OSF.IO/ABTJU.

[155] A Word on “Descriptive” and “Prescriptive” Defining, Merriam-Webster Dict. (n.d.). https://www.merriam-webster.com/grammar/descriptive-vs-prescriptive-defining-lexicography (accessed June 20, 2024).

[156] 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.

[157] 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.

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

[159] S. Gold, J. Flack, VS-Hub, (2022). https://bit.ly/termhub.

[160] W. Gaver, What should we expect from research through design?, in: Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Association for Computing Machinery, New York, NY, USA, 2012: pp. 937–946. https://doi.org/10.1145/2207676.2208538.

[161] A. Ostropolets, PHOEBE, (2020). https://data.ohdsi.org/PHOEBE/ (accessed September 17, 2021).

[162] M.N. Zozus, L.W. Young, A.E. Simon, M. Garza, L. Lawrence, S.T. Ounpraseuth, M. Bledsoe, S. Newman-Norlund, J.D. Jarvis, M. McNally, K.R. Harris, R. McCulloh, R. Aikman, S. Cox, L. Malloch, A. Walden, J. Snowden, I.M. Chedjieu, C.A. Wicker, L. Atkins, L.A. Devlin, Training as an Intervention to Decrease Medical Record Abstraction Errors Multicenter Studies, Stud. Health Technol. Inform. 257 (2019) 526–539.

[163] W.H. Organization, Others, History of the development of the ICD, World Health Organ. (2006).

[164] N. Dean Beaulieu, A.M. Epstein, National Committee on Quality Assurance health-plan accreditation: predictors, correlates of performance, and market impact, Med Care 40 (2002) 325–337. https://www.ncbi.nlm.nih.gov/pubmed/12021688.