Clinical Concept Value Sets and Interoperability in Health Data Analytics

**Sigfried Gold, MFA, MA[1,2]; Andrea Batch, MA[1]; Robert McClure, MD[3];
Guoqian Jiang, MD, PhD[4,2]; Hadi Kharrazi, MD, PhD[5]; Rishi Saripalle, PhD[6];
Vojtech Huser, MD, PhD[7,2]; Chunhua Weng, PhD[8,2]; Nancy Roderer, MLS[1];
Ana Szarfman, MD[9]; Niklas Elmqvist, PhD[1]; David Gotz, PhD[10]
[1]University of Maryland, College Park; [2]Observational Health Data Sciences and Informatics; [3]MD Partners, Lafayette, CO; [4]Mayo Clinic, Rochester, MN; [5]Johns Hopkins University, Baltimore, MD; [6]Illinois State University; Normal, IL; [7]National Library of Medicine, Bethesda, MD; [8]Columbia University, New York, NY; [9]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD;
[10]University of North Carolina, Chapel Hill, NC;**

Abstract

This paper focuses on *value sets* as an essential component in the health analytics ecosystem. We discuss shared repositories of reusable value sets and offer recommendations for their further development and adoption. In order to motivate these contributions, we explain how value sets fit into specific analytic tasks and the health analytics landscape more broadly; 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; the formidable barriers to value set reuse; and 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 standards, infrastructure, and design recommendations are not systematic or comprehensive but invite further work to support value set reuse for health analytics. *The views represented in the paper do not necessarily represent the views of the institutions or of all the co-authors.*

Introduction

This paper focuses on *value sets*[1–5] as an essential component in the health analytics ecosystem.[6–11] While value sets appear in many contexts and serve many purposes and HL7[*] offers a a general specification,[12] 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 collections 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 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;[13] **(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

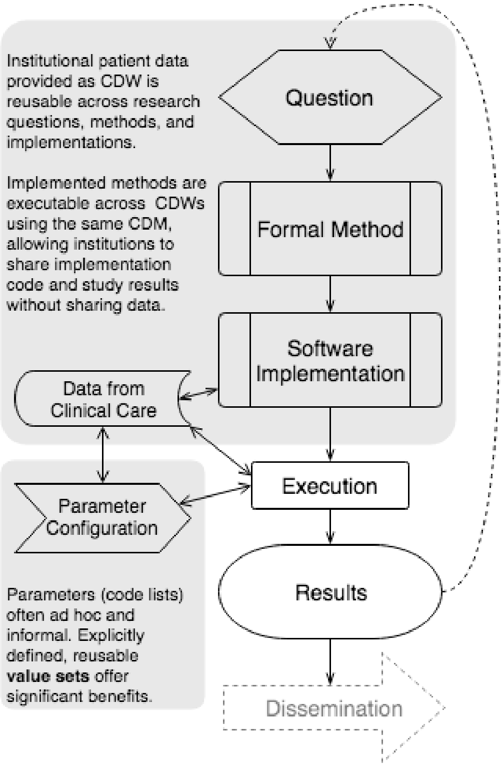
[*] Given the extreme profusion of acronyms used around our topic, we provide meanings and references in an appendix rather at the first use of each acronym.

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.

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, etc.—but makes the discussion relevant to important, non-research secondary uses of health data such as health administration, health economics, public health surveillance, etc., 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.



Institutional patient data provided as CDW is reusable across research questions, methods, and implementations.

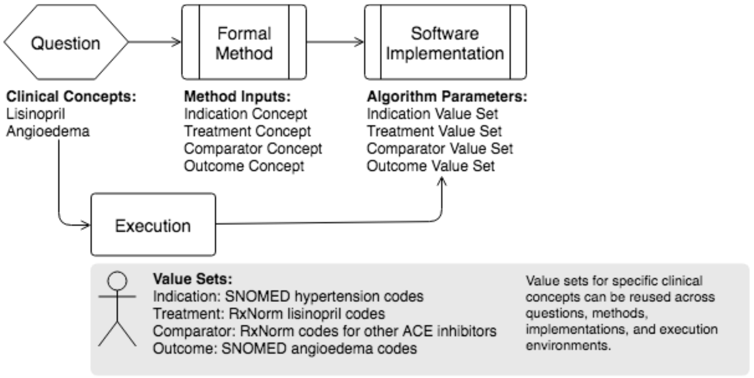
Implemented methods are executable across CDWs using the same CDM, allowing institutions to share implementation code and study results without sharing data.

Parameters (code lists) often ad hoc and informal. Explicitly defined, reusable **value sets** offer significant benefits.

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 statistics, epidemiology, health economics, etc.), 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.



Value sets for specific clinical concepts can be reused across questions, methods, implementations, and execution environments.

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.**

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.[14] 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 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,[15–19] 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, etc.) 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. 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.[8,9,20–29] 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.[30,31]

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,[32] 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, etc. for conditions; NDC, RxNorm, ATC, etc. 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.[33] 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[34] 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[35] 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.

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[13] is a multi-step algorithm making reference to numerous clinical concepts whose definitions are in the form of value sets specified remotely by the US National Library of Medicine (NLM) Value Set Authority Center (VSAC).[4] 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 code 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.

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[36,37] establish norms and language that would suffice if it weren't 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.

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 a specification that identifies a standardized approach to value set metadata and structure: *Characteristics of a Formal Value Set Definition, Release 1.*[12] This specification has been the basis for the 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 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 (eCQMs, etc.), 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 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[38] 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.

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 this 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.

Acknowledgments

We acknowledge and appreciate discussions with and comments from Michael Kahn, Linda Macri, Ben Shneiderman, Catherine Plaisant, Kristin Feeney, Shelley Rusincovitch, Olivier Bodenreider, Daniella Meeker, Chris Chute, Frank DeFalco, Greg Klebanov, Lee Evans, Anthony Sena, Andreas Mathison, Brian Butler, Susan

Winter, Joel Chan, Richard Marciano, David Fram, Patrick Ryan, Jeff Brown, Don Rucker, George Hripcsak, and Jim Cimino. VH work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC).

Appendix: Table of Acronyms

ADE	Adverse Drug Event
CDM	Common Data Model
CDW	Clinical Data Warehouse
CMS	Centers for Medicare and Medicaid Services
DRN	Distributed Research Network
eCQM	electronic Clinical Quality Measure
EHR	Electronic Health Record
ETL	Extract, Transform, Load
HL7	Health Level 7 [cite]
i2b2	Informatics for Integrating Biology and the Bedside[41]
ICD	International Classification of Disease[42]
NCQA	National Committee for Quality Assurance[39]
NDC	National Drug Code[40]
NLM	National Library of Medicine
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PCORNet	National Patient-Centered Clinical Research Network
SME	Subject Matter Expert
SNOMED CT	Systematized Nomenclature of Medicine
UMLS	Unified Medical Language System
VSAC	Value Set Authority Center
VSD	Value Set Definition

References

1. Jiang G, Solbrig HR, Chute CG. Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network. J Biomed Inform. 2011 Dec;44 Suppl 1:S78–85.
2. Jiang G, Solbrig HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. J Am Med Inform Assoc. 2012 Jun;19(e1):e129–36.
3. Peterson KJ, Jiang G, Brue SM, Liu H. Leveraging Terminology Services for Extract-Transform-Load Processes: A User-Centered Approach. AMIA Annu Symp Proc. 2016;2016:1010–9.
4. Bodenreider O, Nguyen D, Chiang P, Chuang P, Madden M, Winnenburg R, et al. The NLM value set authority center. Stud Health Technol Inform. 2013;192:1224.
5. Jiang G, Kiefer R, Prud'hommeaux E, Solbrig HR. Building Interoperable FHIR-Based Vocabulary Mapping Services: A Case Study of OHDSI Vocabularies and Mappings. Stud Health Technol Inform. 2017;245:1327.
6. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. Artif Intell Med. 2016

Jul;71:57–61.

7. Mo H, Jiang G, Pacheco JA, Kiefer R, Rasmussen LV, Pathak J, et al. A Decompositional Approach to Executing Quality Data Model Algorithms on the i2b2 Platform. AMIA Jt Summits Transl Sci Proc. 2016 Jul 20;2016:167–75.
8. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013 Aug;51(8 Suppl 3):S22–9.
9. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing Knowledge Consistently Across Health Systems. Yearb Med Inform. 2017 Aug;26(1):139–47.
10. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc. 2015 Nov;22(6):1220–30.
11. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016 Nov;23(6):1046–52.
12. HL7 Standards Product Brief - HL7 Specification: Characteristics of a Formal Value Set Definition, Release 1 [Internet]. [cited 2018 Mar 8]. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=437
13. Centers for Medicare & Medicaid Services, Office of the National Coordinator for Health Information Technology. Statin Therapy for the Prevention and Treatment of Cardiovascular Disease [Internet]. eCQI Resource Center. 2017 [cited 2018 Mar 3]. Available from: https://ecqi.healthit.gov/ecqm/measures/cms347v1
14. Winnenburg R, Bodenreider O. Metrics for assessing the quality of value sets in clinical quality measures. AMIA Annu Symp Proc. 2013 Nov 16;2013:1497–505.
15. Goss FR, Zhou L, Plasek JM, Broverman C, Robinson G, Middleton B, et al. Evaluating standard terminologies for encoding allergy information. J Am Med Inform Assoc. 2013 Sep;20(5):969–79.
16. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc. 2011 May 1;18(3):341–6.
17. Rector AL, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. Appl Ontol. 2009;4(1):51–69.
18. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, McClay J, et al. The SAGE Guideline Model: achievements and overview. J Am Med Inform Assoc. 2007;14(5):589–98.
19. Rector AL. What's in a code? Towards a formal account of the relation of ontologies and coding systems. Stud Health Technol Inform. 2007;129(Pt 1):730–4.
20. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010 Nov;17(6):652–62.
21. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan;19(1):54–60.
22. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016 Dec;64:333–41.
23. Kahn MG. 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 Sciences Library Photograph Collection and Special Collections, University of Colorado Anschutz Medical Campus, Health Sciences Library; Series V: School of Medicine Publications [Internet]. 2007; Available from: https://dspace.library.colostate.edu/handle/10968/737
24. Kuehn BM. FDA's Foray Into Big Data Still Maturing. JAMA. 2016 May 10;315(18):1934–6.
25. Velentgas P, Bohn RL, Brown JS, Chan KA, Gladowski P, Holick CN, et al. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. Pharmacoepidemiol Drug Saf. 2008 Dec;17(12):1226–34.
26. Brown JS, Kulldorff M, Chan KA, Davis RL, Graham D, Pettus PT, et al. Early detection of adverse drug events within population-based health networks: application of sequential testing methods. Pharmacoepidemiol Drug Saf. 2007 Dec;16(12):1275–84.
27. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010 Aug;19(8):858–68.
28. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of

three public repositories. AMIA Annu Symp Proc. 2013 Nov 16;2013:648–56.

29. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600–6.
30. Bakken S. An informatics infrastructure is essential for evidence-based practice. J Am Med Inform Assoc. 2001 May;8(3):199–201.
31. Bowker GC, Star SL. Building information infrastructures for social worlds—The role of classifications and standards. In: Community computing and support systems. Springer; 1998. p. 231–48.
32. Becnel LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, et al. BRIDG: a domain information model for translational and clinical protocol-driven research. J Am Med Inform Assoc. 2017 Sep 1;24(5):882–90.
33. DeFalco FJ, Ryan PB, Soledad Cepeda M. Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure. Health Serv Outcomes Res Methodol. 2013 Mar 1;13(1):58–67.
34. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf. 2012 Jan;21 Suppl 1:23–31.
35. DeFalco F. OHDSI Architecture Workgroup [Internet]. Observational Health Data Science and Informatics Wiki. [cited 2018 Mar 3]. Available from: http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:architecture_wg
36. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4-5):394–403.
37. Cimino JJ. In defense of the Desiderata. J Biomed Inform. 2006 Jun;39(3):299–306.
38. Lanier J. Who Owns the Future? Simon and Schuster; 2014. 411 p.
39. Dean Beaulieu N, Epstein AM. National Committee on Quality Assurance health-plan accreditation: predictors, correlates of performance, and market impact. Med Care. 2002 Apr;40(4):325–37.
40. National Drug Code Directory [Internet]. U.S. Food & Drug Administration. [cited 2018 Jun 27]. Available from: https://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm
41. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar;17(2):124–30.
42. Organization WH, Others. History of the development of the ICD. World Health Organization. 2006;