# Infrastructures and Interfaces to Encourage Value Set Reuse for 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**

*Interoperability in health analytics—the development of semantically interoperable electronic phenotypes, cohort definitions, quality measures and other algorithmic objects—depends more on the development, refinement, and reuse of high-quality value sets than has been previously recognized. There is widespread interest in such health analytics objects, which depend on clear definitions and computable specifications of clinical concepts; for data coded using controlled vocabularies, clinical concepts are specified in the form of value sets, i.e., lists of vocabulary codes. Our framework for describing and addressing value set management raises the issue of overlapping vocabularies and builds on Cimino's desiderata for controlled vocabularies with the idea of a concept-agnostic orientation to terminology resources. We gather, supplement, and present recommendations for the design of standards, infrastructures, and interfaces to support value set reuse.*

The views represented in the paper do not necessarily represent the views of the institutions.

**Introduction**

Interoperability in health analytics—the development of semantically interoperable electronic phenotypes, cohort definitions, quality measures, and other health analytics objects (HAOs)—depends more on the development, refinement, and reuse of high-quality *value sets* than has been previously recognized. Widespread interest in HAOs and platforms for using them is evident in the informatics literature.[1–6] Algorithmic HAOs depend on the definition and computable specification of clinical concepts. For data encoded using controlled vocabularies, clinical concepts are specified in the form of value sets, i.e., lists of vocabulary codes.

Systematic reuse with audit trail capabilities and ability for automation are critical to the impact of analytics technologies; their benefits—replicable study protocols, comparable results, the evolution and dissemination of best practices, etc.—arise as semantically interoperable HAOs are created, agreed upon by stakeholders, shared, refined, and reused. Real-world reuse faces formidable challenges in developing tools that are technically interoperable across contexts, and much greater challenges in persuading diverse organizations and communities to adopt them.

Our aim in this paper is to formulate a set of infrastructure and interface recommendations for the design of value set management platforms. Satisfaction of issues raised would increase the quality of value sets through improved understanding of value set content, thereby clarifying benefits of value set sharing while decreasing the burden of reuse. To that end, we offer the following contributions: (1) a brief survey of the health analytics landscape informing our vision of value set reuse and their importance to the development of interoperable electronic phenotypes and other HAOs; (2) an approach to problems that arise when defining concepts in the context of multiple, overlapping controlled terminologies; (3) recommendations for value set management technologies; and (4) suggestions for future work, including immediate and practical steps to move toward value set reuse.

Submitted to the American Medical Informatics Association for the AMIA 2018 Annual Symposium.

## Background

This paper focuses on value sets for several reasons: they are essential components of many other types of HAOs; they are simpler than algorithmic HAOs such as electronic phenotypes and quality measures insofar as they can be expressed as enumerated code lists, i.e. static data objects that don't require specific programming language syntax or execution; their development and curation can be managed (somewhat) independently from the objects that depend on them; and established standards, platforms, 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 straightforward, 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, consider Organization A and Organization B belonging to a distributed research network (DRN). Org. A defines statin medications as a particular list of RxNorm or NDC codes for use in a cohort study. With semantic interoperability, Org. B can execute the study in their own environment with comparable results. *Reuse* would mean, e.g., that if Org. B wanted to design a different study relying on a definition of statin medications, they would insert the existing definition rather than defining it again. This requires: 1) that they can find it; 2) that they do find it (implying that searching for it is immediately preferable to defining it anew); 3) they can verify that it serves their current purpose; 4) if it doesn't quite, they modify it so it does and document their change in an easily auditable way so potential future users will understand the difference and, in turn, use or modify the version closest to their own needs.

### *Barriers Against Reuse*

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. Code selection is likely to constitute a small fraction of the analytic work and, perhaps, not the most interesting fraction. The logical complexities involved in crafting cohort definitions and other analytics present a good supply of technical and cognitive challenges. Value set creation may be a tempting place to find shortcuts.

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 are immediately present in workflows requiring value sets. In our experience, they generally overwhelm the benefits of reuse practices, which would only appear if the analyst could readily identify a well validated, well documented value set relevant to her needs, or in some indefinite future.

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[7] 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).[8] 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.

### *Common Data Models (CDM)*

The emergence of common data models (CDM) over the past decade has made interoperable analytics possible in the drug safety and clinical research communities that have adopted them.[3,4,9–18] 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.[19,20]

Those of us active in the Observational Health Data Sciences and Informatics (OHDSI) community have a particular interest in and perspective on value sets (called "concept sets" in that community). OHDSI's rapid growth—in user base, user diversity, and technical platform—has led its Architecture Workgroup[21] 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.

*Definitions*

The following terms will be familiar to many readers. We provide these definitions to point out ways our usage is constrained or colored by our context and perspective, and also to give a scaffolding in which to place brief notes about organizations and projects we refer to.

1. *Controlled vocabulary*. Terminology, ontology, nomenclature, code system, e.g., SNOMED CT, ICD9, ICD10, Read, MeSH, MedDRA, NDC, RxNorm, ATC, NDF-RT, GPI, etc. A lexicon designed and maintained by some authority at a local, organizational level, or, in the case of standard vocabularies, by some cross- organizational standards body. A controlled vocabulary will generally contain an enumeration of concepts, a way of specifying preferred terms and concept synonyms, a definition and unique identifier for each concept, and sometimes affordances for specifying relationships between terms (e.g., "is-a", "caused by", "has anatomical site", etc.)
2. *Encoded health data*. Clinically meaningful data collected or generated in the course of providing care to a patient, including patient demographics, visit records, diagnoses, pharmacy orders, procedures, etc. We limit our discussion to the use of structured and *encoded* data, in which observations and events are referenced in patients' records using codes from controlled vocabularies to represent clinical concepts. Although crucial clinical information exists as narrative notes, medical images, device readings, or lab results, our investigation ignores the interoperability hurdles involved in harmonizing values recorded with different units, or extracting meaning from notes or images, in order to focus on problems involved in the use of encoded terms in analytic contexts.
3. *Secondary use health data*. Collections of health data from multiple patients for purposes beyond providing care or performing administrative tasks related to a single patient, generally some form of research or other analysis.
4. *Health data analysis.* The application of exploratory or evaluative statistics, visualization, or machine learning techniques to health data to gain insights or answer questions in contexts such as quality of care or cost-effectiveness analysis or observational research. For example, what is a 20-year mortality rate of patients with newly diagnosed hemophilia B? Or, is raloxifene better than alendronate in treating osteoporosis?
5. *Health analytics object (HAO)*. A blanket term and acronym we've coined to cover the wide array of standards, algorithms, and semantic resources used in health data analysis. Sometimes it is useful to refer to them all together, but more often an explicit or implied qualifier indicates a subset of them:
    5.1. *Executable or algorithmic* HAOs specify processes or computations, not static data. Electronic phenotypes, cohort definitions, clinical quality measures (eCQM), data quality measures (DQM), population statistics, observational study protocols, etc. are examples of algorithmic HAOs; controlled vocabularies, data models, value sets, etc. are examples of non-executable HAOs.
    5.2. *Computable* HAOs are formally defined with a precise syntax allowing then to be used in software algorithms. Electronic clinical quality measures (eCQM),[7] for instance, have both computable and

uncomputable aspects: their XML renderings are made for use in algorithms and are hence computable. But the algorithms they describe for calculating standard measures of clinical quality must be implemented by users to function in specific contexts in order to execute. In this way they are executable but not computable.

5.3. *Reusable or interoperable*. We apply the term "health data *objects*" to various things we use or create in the course of health data analysis because they *could* be reusable, though often they are not. A method for finding the incidence rate of myocardial infarction could be implemented without reusability as a SQL query written for a local data warehouse; or it could be structured for reuse according to certain specifications for a particular CDM platform and become one object among many in a repository of population statistics measures.

6. *Common data model (CDM)*. CDMs are each centered around a specific data model, but the term is often used as a synecdoche, referring to the whole system of software, infrastructure, organizations, and DRNs based on that model. CDMs allow clinical data networks to share queries, observational study methods, and analytic code. They facilitate the use and reuse of computable and executable HAOs. 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.

6.1. *Observational Health Data Sciences and Informatics (OHDSI).*[22–24] OHDSI has been established as a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics. OHDSI is the community, software, and infrastructure surrounding the Observational Medical Outcomes Partnership (OMOP) data model. Executing DRN studies is straightforward but requires explicit cooperation by partners. Provides a large open source repository of code including ETL, data quality, population measures, data retrieval API, web-based analysis interface, R methods library, electronic phenotypes, cohort definitions. It has an active, growing ecosystem of academic medical centers, pharmaceutical and insurance companies, data and business intelligence vendors, and regulators. OHDSI has been unusual in attracting an active, diverse developer community around its open source platform. Tensions are present, but cooperation is the norm. Other CDMs have larger user communities and extensible, open source platforms, but software development is generally confined to a single organization.

6.2. *i2b2*.[25,26] Extensible, open source analytics platform with a very abstract data model. Mostly used for research or cohort selection from local patient data warehouses. Requires detail coordination of ETL procedures and semantic resources for use in DRNs.

6.3. *Sentinel (Mini-Sentinel)*. FDA-launched initiative for active surveillance. Central control of federated DRN. Central authority dictates data model, vocabulary structures and distributes data quality and research queries to data owners. Data model very similar to OHDSI/OMOP.

6.4. *PCORNet*. Focused on clinical research with patient-reported outcomes (PRO) rather than drug safety. Data model based on Sentinel, but intended to also capture PROs, which are not yet represented well in standard controlled vocabularies.

7. *Value set*. Code set, code list, concept set. 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.

7.1. *In the context of primary use health data*. The permissible values a data element can take; a list of related terms (and associated codes) with *different* meanings, from which (usually) a single term should be chosen to represent a specific concept in the context of the associated data element the value set is used by. Value sets created for primary use are used for data capture (i.e.: a drop-down list,) not data analysis, and therefore are rarely used to identify a specific patient cohort. Although this is not the type of value set we discuss, we have found it important to distinguish and clarify the term's different meanings to prevent miscommunication.

7.2. *In the context of secondary use health data*. A group of codes from one or more controlled vocabularies, selected at any appropriate level of granularity, that can be used to specify a patient cohort, a health outcome of interest or a risk factor. Value sets can be defined by simple enumeration, by using term relationships within or across code systems, or by combining or modifying existing value sets. In this context, the terms in the value set are considered equivalent vis-à-vis the question being asked. A value set for cardiovascular illness might include terms for hypertension, heart failure, and numerous other conditions; with all patient events matching this value set, diverse as it is, considered as instances of cardiovascular illness.

7.3. *In the context of common data elements (CDEs)*. An emerging approach to value sets not addressed here but important to note in a current treatment of the subject. Standardized value set content can be realized through the perspective of meta-data management and registry. One such effort is the ISO/IEC 11179 standard, which specifies a meta-data model for representing the common data elements (CDEs). Enumerated value domains are composed of one or more 'permissible values', each of which represents a valid value for the field. Each of these values, in turn, is tied to a corresponding 'value meaning', which represents the intended meaning of the permissible value in the context of the value domain. The National Cancer Institute (NCI) has implemented the ISO/IEC 11179 standard in the Cancer Data Standards Repository (caDSR) for cancer studies. Recently, a NIH CDM portal has been established for facilitate the use of CDEs in a variety of clinical research studies[27].

8. *Terminology and value set management*. Two overall approaches with very different architectures are prominent in terminology management, services-based and integration of data and vocabularies in the same CDM schemas. Each have significant advantages. Terminology services and value set standards preceded the emergence of CDMs and CDM platforms have largely ignored them as they introduce a level of indirection and external dependencies that complicate the already complicated problems of governing vocabulary inclusion and supporting performant use of semantic resources in processing HAOs.

8.1. *Services-based standards*. Value set management solutions built on terminology services standards access vocabularies through URIs pointing to remote, authoritative locations where the vocabularies are published by the organizations responsible for them.

8.1.1. *Common Terminology Services (CTS2)*. An Object Management Group specification for managing code systems (called vocabularies in this document), code system versions, value sets, and value set definitions.

8.1.2. *The Value Set Authoring Center (VSAC)*. An NLM-managed platform for authoring, validating , maintaining, sharing, and distributing value sets.

8.1.3. *Fast Health Interoperability Resources (FHIR)*. A next generation clinical data standards framework developed by Health Level 7 (HL7.) The FHIR specification contains a terminology module that provides a collection of terminology resources.

8.2. *Integrated with patient data*. In CDM platforms, vocabularies are copied and merged into a small set of tables directly available to queries executed on the patient data repositories. Governing the set of vocabularies and vocabulary versions supported by a CDM platform presents particular problems. CDMs use one of three governance models in determining the vocabularies they support—which constrains the terms available for queries and limits clinical data to sources using those terminologies.

8.2.1. *Centralized authority*. Sentinel, PCORNet.

8.2.2. *Complete user discretion*. i2b2 only specifies the structure of terminology data. For sharing of code and queries across data sources to be semantically meaningful, it is up to users to synchronize vocabulary data.

8.2.3. *Hybrid*. OHDSI allows use of a wide, cross-domain, international range of "source vocabularies" and a centrally authorized set of "standard vocabularies." When clinical data is imported into the CDM, original source codes are retained. If a code does not come from an authorized standard vocabulary, it is mapped to a code that does.

*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[28,29] establish norms and language that would suffice if it weren't for the need to consider value sets that draw from overlapping vocabularies. The "concept orientation" desideratum calls for the concepts in a vocabulary to be nonvague, nonambiguous, and nonredundant. OHDSI's OMOP vocabulary system accomplishes concept orientation by singling out certain concepts (or whole vocabularies) as "standard". But OMOP's collection of vocabularies can also be considered as an undifferentiated heap of concept-agnostic terms, leaving concept orientation as an exercise for the user. To some degree this is the approach taken in DeFalco, et al., "Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure."[30] They take terms from three different drug classification vocabularies (ATC, NDF-RT, and ETC) and follow mappings to three overlapping sets of NDC codes, which they combine into the value set they use to represent opioid exposure.

A concept-oriented vocabulary or ontology divides the relevant universe into a set of nonvague, nonambiguous, and nonredundant concepts. Concepts are the fundamental units of meaning in a vocabulary, distinct from terms, labels, or synonyms, which are names used to denote these concepts, to convey their meaning. A concept-agnostic orientation, on the other hand, makes no judgement as to the success of any vocabulary in achieving concept orientation. It makes no distinction between terms and concepts because concept orientation is lost when overlapping vocabularies are combined (putting aside the ways it can be restored in systems like OHDSI or UMLS). Where a concept-oriented vocabulary calls one object a concept and another a term, a concept-agnostic orientation calls them all terms. Designers of a value set are left to draw on whatever terms and mappings their terminology system contains to collect the set of terms (or term-generating rules) that best reflect their intended concept as coded in the data they will query.

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 will reflect context or specific coding practices in the data, 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.

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.

**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, not merely a technical possibility.

***Value set specifications***. CTS2 and FHIR already provide standards compatible with many of the following recommendations. OHDSI, as mentioned above, is developing its own OpenAPI REST specifications. Possible avenues for achieving a set of common or harmonized specifications are described below. 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*.[31] This specification has been the basis for the FHIR value set resource.

*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 cardiac arrhythmia 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, 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; or for *provenance documentation*. For various reasons, the designer of a value may want to make reference to other value sets without executing their definitions in their own processing. These references could consider other value sets as parents, siblings, or of general interest.

***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, perhaps, hide license-protected codes from unlicensed users, compromising use of value sets that include them, but allowing full functionality for value sets using only unrestricted codes.

*Open, public, crowdsourced curation with AI support*. 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 some user-chosen authority.

*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 to satisfy them all. Rather, they are meant as 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*. Current or potential projects we are aware of include:

- *OHDSI—FHIR*. Two of us are involved in OHDSI's FHIR Working Group (GJ as chair, SG as participant), creating data model mappings to allow interoperability across these two technologies. The group has not yet addressed semantic interoperability.
- *VSAC—OHDSI*. Build bridge or interface to put CTS2 API in front of OMOP vocabulary collections to take advantage of cross vocabulary indexing, etc. Externalize value set management for OHDSI through a CTS2 API that shares storage with VSAC (https://github.com/cts2/vsmc-service). In this way, the value sets from both platforms can be accessed and potentially harmonized through a standard CTS2 API.
- *UMLS—OHDSI.* Concept groupings and mappings in the OHDSI vocabulary CDM keep being refined through community-based feedback and mechanisms may be established to identify well-tested groupings and mappings for refining the UMLS for quality assurance.

- *Cross-CDM*. Each CDM community has its own problems to address and its own approach to vocabulary management. There are efforts underway to harmonize or build bridges between CDMs and address their semantic interoperability issues, but this does not appear to be a priority for these communities.

*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 HAOs. Jaron Lanier[32] describes how companies scramble for the winner-take-all spoils of controlling "siren servers", central hubs for the sharing of crowd-sourced data. Free and open source software (FOSS) around CDMs can address this problem for platform code. Blockchain or similar technology could be explored as a way of consolidating repositories without granting control to a single authority.

***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. The demands of such interfaces could be a tall order for FOSS developers hoping to compete with commercial vendors, but our own experience and ongoing work in this area suggests that practical innovation can emerge from combinations of FOSS, academic research, and participation from enlightened health sciences and services vendors and government agencies. OHDSI operates on this model. Commercial vendors are able to provide proprietary platforms leveraging the OMOP CDM and OHDSI tools and methods, but the community-fostered FOSS resources are of high enough quality to serve a diverse range of several well-financed organizations, who in turn extend these resources towards the general benefit.

*Modularize 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.

*Visualize semantic graph*. Designing an interface for semantic exploration, understanding, and navigation is challenging with some individual vocabularies (e.g., ontologies like SNOMED), 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 methods and results, such as they are, are a potpourri of software engineering, **HCI,** information studies, sociology of technology and organizations, philosophy, journalism, instruction, and polemics. Compared to other papers on value sets[8,24,33–35] this paper is informal. We suggest more formal approaches in the future work section below, but our initial effort to concisely frame and approach the problem seemed to require this broad treatment, solutions that encourage and effectively support value set reuse—and HAO interoperability generally—will require contributions from a broad range of methods and disciplines and perspectives.

Insofar as the definitions section give a picture of practices surrounding secondary use of health data, it 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.

**Future work**

More research on value sets and reuse of health analytics objects is needed. A thorough environmental scan of medical terminology services as they relate to the interoperability and reuse of semantic resources would be of great value.

As the technology and standards for value set management can be functionally separated from controlled vocabularies on the one hand and from algorithmic analytics development on the other, we recommend convening a working group of cross-domain stakeholders and experts to consolidate and disseminate research and practical efforts around value sets and multi-vocabulary semantic interoperability. This could be an appropriate venue to explore the technical and social feasibility of developing open, trusted semantic resource repositories with technologies like blockchain.

Continued work on visualization interfaces for semantic graphs is needed, as is research on design methodologies for large, multi-faceted problems like value set reuse.

**Conclusion**

We have framed the problem of real world reuse of value set and presented a case for it being considered of critical importance for the development of interoperable electronic phenotypes, cohort definitions, and other resources for health analytics. We identify barriers and benefits to cooperative solution building across terminology services and CDM communities. Our framework for describing and addressing value set management raises the issue of overlapping vocabularies and builds on Cimino's desiderata for controlled vocabularies with the idea of a concept-agnostic orientation to terminology resources. We gather, supplement, and present a collection of desiderata for the design of standards, infrastructures, and interfaces to support value set reuse.

**Acknowledgments**

We acknowledge and appreciate discussions with and comments from Michael Kahn, Linda Macri, 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.

**References**

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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
8. 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.
9. 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.
10. 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.
11. 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.
12. 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
13. Kuehn BM. FDA's Foray Into Big Data Still Maturing. JAMA. 2016 May 10;315(18):1934–6.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. Bakken S. An informatics infrastructure is essential for evidence-based practice. J Am Med Inform Assoc. 2001 May;8(3):199–201.
20. 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.
21. 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
22. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A. 2016 Jul 5;113(27):7329–36.
23. Kury F, Huser V. Converting the data in the US CMS Virtual Research Data Center to the OHDSI Common Data Model version 5. In: OHDSI Symposium, October 2015 [Internet]. 2015. Available from: https://www.researchgate.net/profile/Fabricio_Kury/publication/304251805_Converting_the_data_in_the_US_CMS_Virtual_Research_Data_Center_to_the_OHDSI_Common_Data_Model_version_5/links/576aaa2908aefcf135bd360e.pdf
24. 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.

25. 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.
26. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc. 2016 Sep;23(5):909–15.
27. Common Data Element (CDE) Resource Portal [Internet]. U.S. National Library of Medicine; 2012 [cited 2018 Mar 7]. Available from: https://www.nlm.nih.gov/cde/
28. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4-5):394–403.
29. Cimino JJ. In defense of the Desiderata. J Biomed Inform. 2006 Jun;39(3):299–306.
30. 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.
31. 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
32. Lanier J. Who Owns the Future? Simon and Schuster; 2014. 411 p.
33. 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.
34. 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.
35. 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.
37. Bodenreider O. Experiences in visualizing and navigating biomedical ontologies and knowledge bases. In: Proceedings of the ISMB. mor2.nlm.nih.gov; 2002. p. 29–32.