Identifying Appropriate Reference Data Models for Comparative Effectiveness Research (CER) Studies Based on Data from Clinical Information Systems

Omolola I. Ogunyemi, PhD, Daniella Meeker, PhD,† Hyeon-Eui Kim, RN, MPH, PhD,‡ Naveen Ashish, PhD,§ Seena Farzaneh, BSc,‖ and Aziz Boxwala, MD, PhD‡*

Introduction: The need for a common format for electronic exchange of clinical data prompted federal endorsement of applicable standards. However, despite obvious similarities, a consensus standard has not yet been selected in the comparative effectiveness research (CER) community.

Methods: Using qualitative metrics for data retrieval and information loss across a variety of CER topic areas, we compare several existing models from a representative sample of organizations associated with clinical research: the Observational Medical Outcomes Partnership (OMOP), Biomedical Research Integrated Domain Group, the Clinical Data Interchange Standards Consortium, and the US Food and Drug Administration.

Results: While the models examined captured a majority of the data elements that are useful for CER studies, data elements related to insurance benefit design and plans were most detailed in OMOP's CDM version 4.0. Standardized vocabularies that facilitate semantic interoperability were included in the OMOP and US Food and Drug Administration Mini-Sentinel data models, but are left to the discretion of the end-user in Biomedical Research Integrated Domain Group and Analysis Data Model, limiting reuse opportunities. Among the challenges we encountered was the need to model data specific to a local setting. This was handled by extending the standard data models.

Discussion: We found that the Common Data Model from the OMOP met the broadest complement of CER objectives. Minimal information loss occurred in mapping data from institution-specific data warehouses onto the data models from the standards we assessed. However, to support certain scenarios, we found a need to enhance existing data dictionaries with local, institution-specific information.

Key Words: clinical informatics, data modeling, common data models, comparative effectiveness research, semantic interoperability, syntactic interoperability

(*Med Care* 2013;51: S45–S52)

From the *Center for Biomedical Informatics, Charles Drew University of Medicine and Science, Los Angeles; †RAND Corporation, Santa Monica; ‡Division of Biomedical Informatics, University of California, San Diego, La Jolla; §Calit2, University of California, Irvine, Irvine; and ‖Bioinformatics/Medical Informatics, San Diego State University, San Diego, CA.

Some of the data in the paper was presented at the July 2012 EDM Forum meeting in Orlando, Florida.

Supported in part by the Agency for Healthcare Research and Quality (AHRQ). This paper was funded by a contract from Academy Health. Paper authors were also supported by the following Grant: AHRQ 1R01HS19913 (SCAlable National Network for Effectiveness Research).

The authors declare no conflict of interest.

Reprints: Omolola I. Ogunyemi, PhD, Center for Biomedical Informatics, Charles Drew University of Medicine and Science, 2594 Industry Way, Lynwood, CA 90262. E-mail: lolaogunyemi@cdrewu.edu.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.lww-medicalcare.com.

ISSN: 0025-7079/13/5108-0S45

OVERVIEW

Comparative effectiveness research (CER) seeks to answer questions about the impact of an intervention, treatment, or exposure on outcomes or effectiveness by conducting secondary analyses of data collected during normal course of health care.[1,2] It therefore frequently relies upon data from sources such as electronic health record (EHR) systems and administrative claims databases. Our definition of CER encompasses treatments, interventions, or exposures directed at patients and their associated outcomes, as well as interventions directed at health care providers and the effect of these interventions on patient outcomes. Comparative effectiveness studies offer great potential for valuable insights about reducing health care cost, improving health policy decisions, and advancing health care–related research.

Although randomized clinical trials remain the gold standard for assessing the impact of treatments or interventions, data sources that capture routine clinical practice can provide a wealth of information on treatment and intervention outcomes that might be difficult to ascertain from a randomized clinical trial because of trial design complexity, high costs, and other factors. Data from health care information systems collected during the course of normal care have the advantages of providing: (1) an accurate picture of the health care services actually provided in different care settings; (2) greater numbers; and, (3) more diverse patient populations. These electronic data enable researchers to assess the impact of real-life clinical practice on patient outcomes. However, reaping the full benefits of vast quantities

of accumulated data requires overcoming many challenges, one of which is to make data collected and stored in different systems and locations interoperable.[3–7] Lack of an open and shared information infrastructure impedes analysis of data from disparate sources.[8]

Consider software created to summarize blood pressure information collected from 2 different clinical practices—the summarizing software must be able to recognize how to read data files into memory and subsequently recognize which parts of the different files contain blood pressure information. If one practice reports blood pressure as "high," "normal," or "low," whereas the other reports blood pressure as a systolic/diastolic value, custom programming is required to summarize the combined data in a meaningful way. Enabling a larger scale analysis of heterogeneous data requires that ≥ 2 systems involved in collaborative analysis can exchange data or information with each other (syntactic interoperability) and that the different systems involved in data exchange can *understand* and use the exchanged data and information (semantic interoperability). Syntactic interoperability requires making formats for data and messages consistent across the different systems involved in data exchange. Semantic interoperability requires that the meaning of the data is unambiguous and correctly interpreted by both humans and computers that use the data.[9] To achieve semantic interoperability, data from various sources are mapped to standardized terminology systems and annotated with additional information, called metadata, that is critical for correctly understanding the data's meaning. Metadata include key contextual information, for example, where blood pressure readings were taken (emergency room or outpatient setting).

A *data model* specifies a system for representing data elements, metadata, and relationships between different data elements in a specified domain, for example, clinical CER. When used as the design of a database, this set of structural specifications are referred to as a "schema." To address both the syntactic and the semantic interoperability issues that arise when attempting to utilize data that is idiosyncratically generated from various sources, typically, multisite projects adopt a common data model that meets specific purposes of the project.[10–16] As common data model creation or selection is often driven by a specific purpose, no existing common data models are robust enough to encompass every domain or data representation need. However, data collected during the normal course of health care are sufficiently constrained that a common standard for research purposes can address a wide variety of research questions.

The approach of standardizing data interchange is consistent with the findings of other researchers: use of a common data model promotes analytic consistency across sites and ensures that the results from similar analyses at different sites are comparable and not affected by potentially different protocol interpretations.[16]

In this paper, we examine the challenges associated with representing and mapping data for analyses in CER studies that use data taken from multiple EHRs and associated data warehouses. We outline a rationale for adopting a common or reference data model and compare the strengths and weaknesses of existing data models that can be used as a common or reference data model. We assess the impact of having a common data model on the approach to data collection and exchange, and present lessons learned. Using a finite set of data elements related to CER drawn from an actively used research data warehouse, we also present an evaluation of the modeling challenges and data or information loss that can occur when using different existing data models.

A glossary of the technical terms and acronyms used in the paper is presented in Table 1.

COMMON CHARACTERISTICS OF CER STUDIES

CER is characterized by research and research infrastructure development to improve medical decisions and clinical outcomes by comparing various drugs, treatments, and other interventions.[17] Common epidemiological designs include case-control, parallel cohort, and self-controlled case series comparing outcomes of alternative treatments. In order for causal inference to be sufficiently robust in nonrandomized studies, analytic models require collection of substantial additional data and metadata beyond the treatments and outcomes.[18–20] Typically, methods relying upon factors that are correlated with treatment selection but not with outcomes require data not found in minimal data models, for example, insurance coverage, policy context, and location details. Comprehensive data models that accommodate rich metadata augment the validity of observational analyses.

CER, and more recently, patient-centered outcomes research, often include analyses of outcomes that are not strictly clinical, such as outcomes related to utilization, expenditures, and quality of life.[21] This may require representation in data models of observations of patient-reported outcomes that are not typically collected in the course of routine clinical care, although required by the Centers for Medicare and Medicaid Services in certain situations, for example, care received in nursing facilities.[22] These common characteristics of CER studies drive what is needed from a data model at a minimum: representation of patient demographics, drugs, procedures, outcomes or observations, providers, health care facilities, insurance features, and payments.

RATIONALE FOR ADOPTING A COMMON/ REFERENCE DATA MODEL

Adopting a common or reference data model lays the groundwork for achieving syntactic and semantic interoperability so that comparable CER analyses can be performed across research study sites. This section examines some benefits and potential limitations of adopting a common data model.

Benefits of Using a Common Data Model

Schema Standardization

Using a common data model has an important practical value of providing a "checklist" of required and optional data elements. Required data elements are typically identifiers for entities represented in the data (eg, patient, encounter), and content that is necessary for linking data across multiple entities.

TABLE 1. Glossary of Terms and Acronyms Used

Terms Used	Definition
Data element	An atomic or indivisible unit of data
Data model	An abstract model that describes the ways in which data collected for business purposes (eg, clinical practice, medical claims) should be structured and organized, including a specification of the data elements that make up the model
Data mapping	The process of identifying relationships among and linking the data elements present in 2 distinct data models
Data dictionary	A centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format
(Database) schema	The formal language description of the structure of the data in a database management system
Database query	Specially written software code sent to a database in order to get information back from the database
Data warehouse	A central repository of data which is created by integrating data from one or more disparate sources (eg, several different EHR and research data systems)
(Data model) heterogeneity	Refers to different ways of representing and storing the same data. Also referred to as *schematic* heterogeneity
Foreign key	A field in a database table that refers to the primary key of another database table (see definition of primary key below)
Interoperability	The ability of diverse systems and organizations to work together (interoperate)
(Information) mediator	A software system that provides unified access to data in distributed, and possibly heterogeneous data repositories
Metadata	Is a description of the data (as in a database), and is often described as "the data about the data"
Overgeneralization	Refers to semantic details that are lost when a local data model is mapped to a reference data model because the reference model does not provide ways to represent data at a sufficient level of detail. For example, if a reference model provides a single field for blood pressure without additional fields that capture the method and body site of the measurement, different types of blood pressure measurements from a local dataset will be mapped to the single field "blood pressure" in the reference model without additional details, resulting in overgeneralization
Primary key	A field (or group of fields) that serves as a unique identifier within a database table, eg, the field "social security number" in a database table identifying details about a person
Syntactic interoperability	The capability of ≥ 2 systems to communicate and exchange data
Semantic interoperability	The ability (of ≥ 2computer systems) to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems.
(Semantic) concept	An idea or thought that corresponds to some distinct entity or class of entities, or to its essential features, or determines the application of a term, and thus plays a part in the use of reason or language
Semantic link	An explicit relationship between 2 (semantic) concepts
Terminology	A set of concepts and relationships used to describe terms and relationships in an application of interest. A fine-grained (more granular) terminology uses more terms and more detail to describe an application area of interest than a coarse-grained (less granular) terminology
Acronym	Definition
NLP	Natural Language Processing, a discipline within computer science, artificial intelligence, and computational linguistics aimed at machine driven understanding of natural language text
HL7	Health Level Seven International; the global authority on standards for interoperability of health information technology
XML	Extensible Markup Language, a "tag-based" language for marking structure in elements in any text data

(*Continued*)

Terminology Standardization

Within a single site or study, a standardized terminology guides data managers and investigators as they make decisions about representing local data using standardized concepts that have been established in the broader literature. Across multiple sites, representing locally generated data with standardized concepts promotes parity in independent analyses and allows data from different sources to be combined into a single coherent dataset.

Potential Limitations of Using a Common Data Model

Unmappable Data

Many EHRs allow providers to include full-text notes regarding a patient's treatment and outcomes. Full text generally can only be mapped to a common terminology by using Natural Language Processing (NLP) technology or through manual coding. However, with NLP, there is a possibility that even if identical algorithms are employed at different sites, context may generate different results. Local customization of NLP algorithm parameters may require generating training data for NLP that is much more costly than using "out of the box" methods that have been trained on external datasets.[23] Manual coding is inefficient and impractical at the scale needed for CER studies.

Information Loss

When a terminology with detailed descriptions (ie, a fine-grained terminology) is mapped into a terminology that has fewer details or is coarse-grained (eg, a mapping from SNOMED to ICD9 or from ICD10 to ICD9), information is lost. It is thus beneficial when possible, to include the terminology source value even after mapping to a common data model. Maintenance of source records also facilitates revisions of terminology mappings if the common model evolves.

METHODS

Analysis of the Strengths and Weaknesses of Existing Data Modeling Standards

We examined the strengths and weaknesses of existing data models with regards to meeting a requirement that the model support a broad range of current and future CER

projects. We examined the Clinical Data Interchange Standards Consortium (CDISC) Analysis Data Model (ADaM); the Biomedical Research Integrated Domain Group (BRIDG) model; the Observational Medical Outcomes Partnership (OMOP) Common Data Model versions 2 and 4; and the US Food and Drug Administration (FDA's) Mini-Sentinel Common Data Model (MSCDM) versions 1.1 and 2.1. A brief overview of each modeling standard follows.

CDISC has developed several standards for exchange, storage, and submission of clinical trials data to the FDA. Their ADaM[14] is a standard for describing the structure, content, and metadata associated with analysis datasets from clinical trials. ADaM has 4 categories of metadata covering the analysis dataset itself as a whole, analysis variables, analysis parameter values, and analysis results. The metadata explain how an ADaM dataset was created from source data.

The BRIDG model aims to represent the semantics of the data from clinical trials and preclinical protocol-driven research.[11] BRIDG is a collaborative effort involving CDISC, the HL7 Regulated Clinical Research Information Management Technical Committee, the National Cancer Institute, and the FDA. The BRIDG model's static semantics are presented through class diagrams and instance diagrams that describe the concepts, attributes, and relationships that occur in clinical and preclinical research. Its dynamic semantics are presented primarily through state transition diagrams that model the behavior of those concepts and how the relationships evolve.

OMOP was set up by the Foundation of the National Institutes of Health to aid in monitoring the use of drugs for safety and effectiveness. OMOP has created common data models that define the primary data elements needed across observational studies. Specifications from OMOP include a data dictionary required for standardizing aggregated data so that such studies can be compared.[13,24] OMOP's goal is to facilitate observational studies that involve using data from different databases, including administrative claims data and EHRs. For our modeling comparison, we examined OMOP Common Data Model versions 2.0 and 4.0 (disclosure: some input from this paper's authors went into the design of version 4.0.).

The Mini-Sentinel is a pilot program sponsored by the FDA. Its goal is to develop comprehensive approaches that facilitate the use of data routinely collected and stored in EHRs for surveillance of marketed medical products' safety. More than 30 academic and private health care institutes are participating in this program as data partners who provide data for the surveillance activity. MSCDM specifies required data content and structure as well as a standardized vocabulary mapping. To achieve semantic interoperability among the disparate datasets, the data partners transform their local datasets into the MSCDM conforming formats.[10,25–28] We analyzed versions 1.1 and 2.1 of the MSCDM.

We compared the different modeling standards based on whether they were easily extensible; adequately captured patient demographic and clinical data; were easy for clinical researchers and data analysts to understand; modeled financial payment and payer data; used standardized vocabularies; modeled insurance plan benefit design and benefit plan data; had widespread real-world usage; had well-defined analytic methods and a user-base for these methods; and had the ability to model nondrug, nonprocedure interventions. As discussed above, these considerations were made in the context of the project's present and long-term goals.

Information Loss and Limitations of Using a Common Data Model

To understand and demonstrate potential limitations and challenges associated with representing local data in a common data model, we mapped the data schema of UC San Diego's Clinical Data Warehouse for Research (CDWR) (Fig. 1) to 4 widely adopted data models: BRIDG, ADaM, OMOP CDM version 4.0, and Mini-Sentinel version 2.1. We were unable to gain access to the detailed database schema for ADaM from CDISC and had to make inferences about primary keys and foreign key relationships across tables. As a consequence, in this paper, we focus on results for the BRIDG, OMOP, and Mini-Sentinel models. We chose these models because their data elements are designed to represent the data from EHRs, clinical data warehouses, or clinical trial management systems. As the BRIDG model does not include a standardized vocabulary, we used the CDWR's local vocabulary for BRIDG in our mapping exercise. We examined how completely and accurately these models represent local data using 2 CER study scenarios (see Supplemental Digital Content for full details, http://links.lww.com/MLR/A507). Our goal was to answer the following questions.

(1) Does every data element have a place in the reference models?
(2) What kinds of extensions or modifications are required to represent the testing scenarios?
(3) Does the model transformation result in overgeneralization?
(4) Does the model transformation result in missing nuance or attribution (eg, reported vs. observed vs. measured), which is critical for data interpretation?
(5) Are there any missing semantic links?

Scenario 1: The first scenario assessed was for a medication therapy management study that involved ascertaining age, sex, race, ethnicity, insurance status, marital status, and other relevant variables for all patients with type-II diabetes seen at several clinics belonging to 1 study site between January 1, 2011 and June 30, 2011. The CER inquiry of interest was as follows: are outcomes different for patients with type-II diabetes whose medication use is managed by *both* a pharmacist and a physician as compared to patients with type 2 diabetes whose medication use is managed *only* by a physician?

Scenario 2: The second scenario assessed was for a local clinical study that involved ascertaining primary care physician name, last visit date, highest hemoglobin A1C level, and other relevant variables for patients between the ages of 25 and 70 who had been seen at a study site's family medicine clinic within the last 18 months.

CDWR

The CDWR consists of 8 major tables that cover entities such as Encounter, Problem, Observation, Patient,

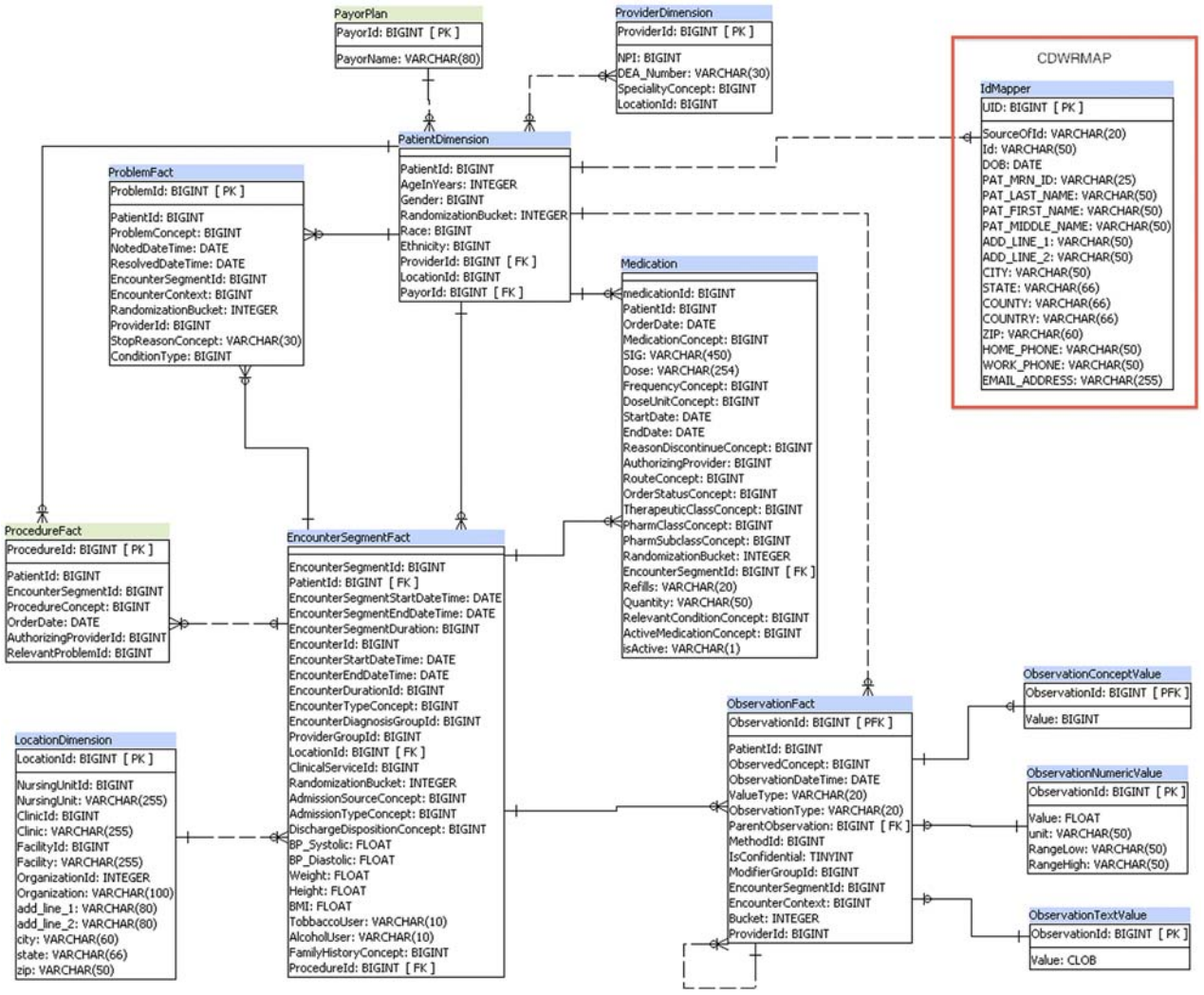

FIGURE 1. Clinical Data Warehouse for Research (CDWR) schema.

Provider, Medication, Procedure, and Location (Fig. 1). Clinical data are extracted, transformed, and loaded (in a process called ETL for short) from UCSD's Epic Clarity EHR system to the CDWR every day.

Model Mapping and Data Querying

One of the authors (S.F.) performed cross-mapping of the CDWR schema to BRIDG, OMOP CDM version 4.0, and Mini-Sentinel. The other authors reviewed the cross-mapping results to ensure accuracy. Queries for the 2 scenarios above were run and the results compared.

Common Data Model Impact on Data Collection, Mediation, and Exchange Approaches

A common data model serves as the reference framework for organizing the data generated from multiple, disparate, and independent study sites, in order to integrate the data for further analyses. A common data model allows 1 query to be specified to obtain data from multiple sites. A typical approach to integrating the data is to aggregate the data from all sites into a single database that uses the schema of the common data model. Data are updated in this database from the sites at regular intervals, for example, every night, once a month. The query is then applied to this database and the results are presented to the user. This single database approach has the disadvantage of data not being current, and sites losing control of their data. A less conventional approach that provides more flexibility is to use an information mediator to query each site. The mediator is a software engine that translates queries composed in the "language" of the common data model to queries to in the "language" of individual data sources, and then transforms the results obtained from various sites into an integrated result in the common data model. Figure 2 presents an example of the architecture that might be associated with data sharing using a mediator.

The transformation from the original clinical data source to a common data model is captured by mediator *rules*. Mediator rules are declarative logical rules that specify

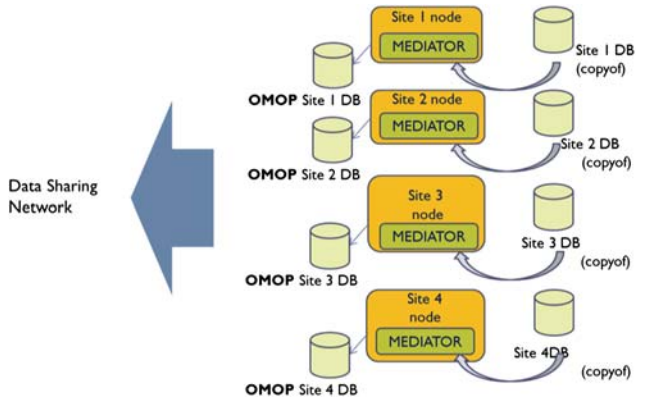

FIGURE 2. Data sharing for analysis example architecture.

how information from the original data source relates to the information in the global (common data) model. The common data model impact on data collection, mediation and exchange approaches section outlines some challenges associated with using a mediator.

RESULTS

Analysis of the Strengths and Weaknesses of Existing Data Modeling Standards

With regards to the process of extending models to meet more CER study needs, we found that for some of the models, formal membership in the sponsor organization is required (including the payment of membership dues in the case of CDISC) in order to be involved in the process of suggesting extensions. The OMOP models were an exception in this regard.

Most of the models adequately captured patient demographic and clinical data, although access to the restricted vocabularies that facilitate translation does require use agreements. The Mini-Sentinel CDM version 1.1 did not model laboratory findings, because the main purpose of this model was to represent events from claims data but not specific clinical data captured in EHRs. However, version 2.1 expands its scope to include laboratory findings. The BRIDG model encompasses the widest range of biomedical fields. However, we found the BRIDG model to be less intuitive for clinical researchers and data analysts to understand than the other modeling standards. This is not surprising, given its role as a domain analysis model for representing a large variety of research study scenarios in detail. OMOP and Mini-Sentinel specified the use of standardized vocabularies, which aid in achieving semantic interoperability, but BRIDG and ADaM did not. With BRIDG, encoding the data with a standardized vocabulary is required. However, BRIDG does not specify the vocabulary systems to use. Instead, it allows users to use standardized vocabulary systems that are compiled through the Enterprise Vocabulary Services of the National Cancer Institute.

OMOP's CDMs, ADaM, and the Mini-Sentinel CDMs had well-defined analytic methods with an existing or growing user-base but BRIDG did not meet this criterion. Most of the data modeling standards examined were designed with a particular use in mind and would require some adaptation for the purposes of our research. OMOP CDM version 4.0, although not perfect, met most of the study's needs with regards to short-term and long-term usage goals. Our findings are summarized in Table 2.

Information Loss and Limitations of Using a Common Data Model

We created and executed queries for the 2 CER study scenarios from the Information loss and limitations of using a common data model section.

OMOP CDM Version 4.0

A majority of the data fields from CDWR were successfully mapped. However, some local extensions to the CDM were required to capture detailed data items that are frequently used for local research projects. Local research use cases often require data at varying levels of granularity: for example, data about provision of care may be needed at a more fine-grained level that specifies details such as "provider name" or at a less fine-grained level that specifies only "specialty care site." Fine-grained data such as "provider name" is unlikely to be shared across independent institutions seeking to use the common data model to facilitate data exchange. This issue was deemed solvable at a local site level by enhancing the local site's dictionary tables. As the data concepts and attributes in this model are appropriately fine-grained, there was no overgeneralization from mapping fine-grained concepts and attributes into coarser-grained ones.

Although we did not find any alterations in semantics when representing the scenarios with the OMOP CDM, the schema has complexities that may result in barriers to adoption. OMOP (and the Mini-Sentinel data models) support both claims and EHR data, but were initially developed with focus on claims data. This results in features that may appear nonintuitive, particularly to users unfamiliar with claims data. For example, claims data identify outpatient office visits as CPT procedure codes, so metadata regarding the provider *with whom* a visit occurred are treated analogously to a provider *by whom* a procedure was conducted.

BRIDG

Nine tables from the BRIDG model successfully covered the current CDWR (see Supplemental Digital Content 1, http://links.lww.com/MLR/A507). The BRIDG model clearly represents the semantics of the data. We did not find any alterations in semantics when representing the scenarios with the BRIDG model (although a similar scenario as with the OMOP CDM in which office visits were represented as procedures was observed).

Unlike OMOP CDM version 4.0, the BRIDG model does not include payer information. We did not observe any overgeneralization in representing the CDWR content with BRIDG. In general, however, BRIDG provides more detailed ways of representing data than the CDWR. For example, (medical) procedure information can be represented using multiple fields such as "methods" and "anatomic site." To simplify matters, highly specialized BRIDG tables and

TABLE 2. Comparison Criteria for Data Models

Data Model	Models Insurance Benefit Design/Plans	Widespread Real-World Usage	Well-defined Analytic Methods and User-Base	Ability to Model Nondrug, Nonprocedure Interventions	
CDISC ADaM	No	Yes	Yes	Limited	
BRIDG	No	Yes	No	Yes	
OMOP version 2.0	No	Yes	Yes	Yes— observation	
OMOP version 4.0	Yes	Yes	Yes	Yes— cohort	
Mini-Sentinel version 1.1	Limited	Yes	Growing	No	
Mini-Sentinel version 2.1	Limited	Yes	Growing	No	
Data Model	**Model is Easily Extensible**	**Adequately Captures Patient Demographic and Clinical Data**	**Easy for Clinical Researchers and Data Analysts to Understand**	**Models Detailed Financial Payments and Plan Data**	**Uses Standardized Vocabularies/Data Dictionaries**
CDISC ADaM	Needs CDISC membership	Partly	Yes	No	No (user defined)
BRIDG	CDISC/HL7/ CaBIG involvement	Yes	No	No	No (user defined)
OMOP version 2.0	Yes	Yes	Yes	No	Yes
OMOP version 4.0	Yes	Yes	Yes	Yes	Yes
Mini-Sentinel version 1.1	No	Laboratory values not modeled	Yes	No	Yes
Mini-Sentinel version 2.1	No	Yes (laboratory values included)	Yes	No	Yes

ADaM indicates Analysis Data Model; BRIDG, Biomedical Research Integrated Domain Group; CDISC, Clinical Data Interchange Standards Consortium; OMOP, Observational Medical Outcomes Partnership.

attribute fields were combined into more general tables and attributes fields.

Mini-Sentinel 2.1

Unlike other models, Mini-Sentinel incorporates standardized concept codes into the tables as separate fields. A few examples of the standardized codes employed by Mini-Sentinel are Logical Observations Identifiers Names and Codes (LOINC) for laboratory and/or clinical observations, and National Drug Codes (NDC) for medication names. Embedding standardized codes into the tables simplifies the data model. However, it also means that users who adopt different coding systems need to map their data to the coding system included in the Mini-Sentinel model in order to use this model to represent their data. For example, the data in the CDWR at the testing site are not encoded with LOINC and NDC, we needed to first map our data to LOINC and NDC before transforming our CDWR model to Mini-Sentinel. Mini-Sentinel does not provide separate tables for providers and service locations. Only the identification codes of providers and service locations are represented in this model. Mini-Sentinel is a relatively compact model with 8 tables that are intended to capture core minimum data. Unlike OMOP, which provides robust ways to represent various clinical observations, Mini-Sentinel specifies a limited number of clinical observations in the laboratory and vitals tables. Therefore, many clinical observations that are not specified as a separate field in these tables cannot be represented. For example, body mass index is a data item required by scenario 2 that cannot be retrieved when the data is represented with Mini-Sentinel.

Common Data Model Impact on Data Collection, Mediation, and Exchange Approaches

There are 2 main challenges we confronted in attempting to use a data mediator.

(1) Building the mediator transformation model: this is an involved task requiring input from data source designers and administrators (ie, those who understand the original clinical data source and its underlying databases quite well). This transformation process is the most expensive part of the mapping process, and is required whether the queries are translated with a mediator, or the source data are transformed into a common model.
(2) Performance: the performance of the mediator is inefficient in situations that involve combining data from >1 large table (ie, tables containing several million records). This is because the mediator is relatively novel technology, and it does not yet have many of the optimization methods that are available to a query written in a language that is native to a database. We expect that this issue will be overcome with further research.

DISCUSSION

We have outlined the importance of syntactic and semantic interoperability in analyzing clinical data from heterogeneous sources for CER. In order to achieve this,

selecting a common data model that can accurately and completely represent these data is the key. Our data modeling comparisons found that although many of the models examined captured a majority of the data elements that are useful for CER studies (patients, drugs, procedures, outcomes/observations, providers, health care facilities, benefit plans, payments, etc.), modeling of insurance plan benefit design and financial plans were most detailed in OMOP CDM version 4.0. We note that standardized vocabularies or data dictionaries were present in the OMOP and Mini-Sentinel data models but would need to be defined by the end-user for BRIDG and ADaM. This is an important issue with regards to achieving semantic interoperability.

Our study on information loss identified a need to extend OMOP's CDM to address local modeling requirements (ie, those not shared for analysis across the network) and nonintuitive ways of modeling office visits in the OMOP and BRIDG data models.

With regards to data mediation, we did not find a mediator robust enough to handle the complex data mapping process from the different clinical information systems present across our study sites. However, more robust mediator software systems are in development, making the query translation option a more viable approach to facilitate data exchange.

ACKNOWLEDGMENT

The authors would like to thank Paulina Paul (UCSD) for her assistance with data modeling and mapping.

REFERENCES

1. Berger ML, Mamdani M, Atkins D, et al. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I. *Value Health*. 2009;12:1044–1052.
2. Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. *Value Health*. 2009;12:1053–1061.
3. Kuchinke W, Ohmann C, Yang Q, et al. Heterogeneity prevails: the state of clinical trial data management in Europe—results of a survey of ECRIN centres. *Trials. [Multicenter Study Research Support, Non-U.S. Gov't]*. 2010;11:79.
4. Lezzoni LI. Multiple chronic conditions and disabilities: implications for health services research and data demands. *Health Serv Res*. 2010;45 (5 Pt 2):1523–1540.
5. Niland JC, Rouse L, Stahl DC. An informatics blueprint for healthcare quality information systems. *J Am Med Inform Assoc*. 2006;13: 402–417.
6. Shortliffe EH, Sondik EJ. The public health informatics infrastructure: anticipating its role in cancer. *Cancer Causes Control*. 2006;17:861–869.
7. Souza T, Kush R, Evans JP. Global clinical data interchange standards are here! *Drug Discov Today*. 2007;12:174–181.
8. Brochhausen M, Spear AD, Cocos C, et al. The ACGT Master Ontology and its applications—towards an ontology-driven cancer research and management system. *J Biomed Inform [Research Support, Non-U.S. Gov't Review]*. 2011;44:8–25.
9. caBIG. caBIG Compatibility Guideline: Achieving Semantic Interoperability; 2008. Available at: https://cabig.nci.nih.gov/community/guidelines_documentation/. Accessed June 1, 2012.
10. Center M-SC. Mini-Sentinel: Common Data Model and Its Guiding Principle; 2010.
11. Fridsma DB, Evans J, Hastak S, et al. The BRIDG project: a technical report. *J Am Med Inform Assoc*. 2008;15:130–137.
12. Network HR. Collaboration Toolkit: a Guide to Multicenter Research in the HMO Research Network; 2011. Available at: http://www.hmoresearchnetwork.org/resources/toolkit/HMORN_CollaborationToolkit.pdf. Accessed June 1, 2012.
13. OMOP. Observational Medical Outcomes Partnership Common Data Model Version 2.0 Specifications; 2011. Available at: http://75.101.131.161/download/loadfile.php?docname=CDM%20Specification%20V2.0. Accessed April 12, 2012.
14. CDISC Analysis Data Model Team. CDISC Analysis Data Model (ADaM) v2.1; 2009. Available at: http://www.cdisc.org/stuff/contentmgr/files/0/854651256c2654c94b03e6da1be6e145/misc/analysis_data_model_v2.1.pdf. Accessed June 22, 2011.
15. Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012;19:54–60.
16. Brown J, Lane K, Moore K, et al. Defining and evaluating possible database models to implement the FDA Sentinel initiative: Harvard Medical School and Harvard Pilgrim Health Care; 2009 Contract No.: HHSF223200831315P. [Report available at http://www.regulations.gov/#!documentDetail;D=FDA-2009-N-0192-0005].
17. Iglehart JK. Prioritizing comparative-effectiveness research—IOM recommendations. *N Engl J Med*. 2009;361:325–328.
18. Leslie RS, Ghomrawi H. *The Use of Propensity Scores and Instrumental Variable Methods to Adjust for Treatment Selection Bias*. Available at: http://www2.sas.com/proceedings/forum2008/366-2008.pdf. Accessed June 1, 2012.
19. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*. 1994;272:859–866.
20. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public health*. 1998;19:17–34.
21. Stewart AL, Ware JE. *Measuring Functioning and Well-being: the Medical Outcomes Study Approach*. Durham, NC: Duke University Press Books; 1992.
22. Mor V. A comprehensive clinical assessment tool to inform policy and practice: applications of the minimum data set. *Med Care*. 2004;42:III50–III59.
23. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17:507–513.
24. OMOP. Observational Medical Outcomes Partnership Common Data Model Version 4.0 Specifications; 2012. Available at: http://75.101.131.161/download/loadfile.php?docname=CDM%20Specification%20V4.0. Accessed June 15, 2012.
25. Center M-SC. Mini-Sentinel: Principles and Policies, 2012. Available at: http://mini-sentinel.org/work_products/About_Us/Mini-Sentinel-Principles-and-Policies.pdf. Accessed May 30, 2013.
26. Core M-SCC-D. Mini-Sentinel: Overview and Description of the Common Data Model v2.1; 2011. Available at: http://www.mini-sentinel.org/work_products/Data_Activities/Mini-Sentinel_Common-Data-Model_v2.1.pdf. Accessed November 4, 2012.
27. Curtis LH, Weiner MG, Boudreau DM, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):23–31.
28. Platt R, Carnahan RM, Brown JS, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):1–8.

