Journal of Biomedical Informatics 64 (2016) 333–341

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Special Communication

Evaluating common data models for use with a longitudinal community registry

Maryam Garza [a], Guilherme Del Fiol [b], Jessica Tenenbaum [c], Anita Walden [a,d], Meredith Nahm Zozus [c,d,*]

[a] Duke Translational Medicine Institute, Duke University, 2424 Erwin Road, Hock Plaza Box 3850, Durham, NC 27705, USA
[b] Department of Biomedical Informatics, University of Utah School of Medicine, 421 Wakara Way, Room: Suite 140, Salt Lake City, UT 84108, USA
[c] Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Suite 1102 Hock Plaza Box 2721, Durham, NC 27705, USA
[d] Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, 501 Jack Stephens Drive, Mail Slot # 782, Little Rock, AR 72205, USA

ARTICLE INFO

Article history:
Received 2 May 2016
Revised 26 October 2016
Accepted 27 October 2016
Available online 29 October 2016

Keywords:
Common data model
Electronic health records
Data model evaluation

ABSTRACT

Objective: To evaluate common data models (CDMs) to determine which is best suited for sharing data from a large, longitudinal, electronic health record (EHR)-based community registry.
Materials and methods: Four CDMs were chosen from models in use for clinical research data: Sentinel v5.0 (referred to as the Mini-Sentinel CDM in previous versions), PCORnet v3.0 (an extension of the Mini-Sentinel CDM), OMOP v5.0, and CDISC SDTM v1.4. Each model was evaluated against 11 criteria adapted from previous research. The criteria fell into six categories: content coverage, integrity, flexibility, ease of querying, standards compatibility, and ease and extent of implementation.
Results: The OMOP CDM accommodated the highest percentage of our data elements (76%), fared well on other requirements, and had broader terminology coverage than the other models. Sentinel and PCORnet fell short in content coverage with 37% and 48% matches respectively. Although SDTM accommodated a significant percentage of data elements (55% true matches), 45% of the data elements mapped to SDTM's extension mechanism, known as Supplemental Qualifiers, increasing the number of joins required to query the data.
Conclusion: The OMOP CDM best met the criteria for supporting data sharing from longitudinal EHR-based studies. Conclusions may differ for other uses and associated data element sets, but the methodology reported here is easily adaptable to common data model evaluation for other uses.

1. Background and significance

1.1. Defining common data models

Common data models (CDMs) are often used in research when there is a need to exchange or share a set of data for some particular use. A data model is a representation of data typically collected about things or events and the relationships between them [1]; a CDM is used to standardize and facilitate the exchange, pooling, sharing, or storing of data from multiple sources. CDMs specify structure, format, and (to varying degrees) content of data to be pooled or shared, and have been used in clinical research since the early days of multicenter registries and data reporting, when minimum data sets were defined and exchanged [2]. Within the last decade, several CDMs have been collaboratively developed and have risen to the level of *de facto* standards for clinical research data. These include the Sentinel CDM [3], the National Patient-Centered Clinical Research Network (PCORnet) CDM [4], the Health Care Systems Research Network (formerly known as the HMO Research Network) Virtual Data Warehouse [5], the Observational Medical Outcomes Partnership (OMOP) CDM [6], and the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) [7,8]. These models span uses across the continuum of clinical research, from clinical trials to observational studies and post-marketing surveillance. With the current national and international emphasis on sharing research data [9–12], CDMs are increasing in importance. Sharing data via a CDM decreases some barriers to reuse of the data, increasing the likelihood that the data will be used to answer additional scientific questions.

1.2. Data model evaluation

CDM quality has been defined as "the totality of features and characteristics of a conceptual model that bear on its abil-

* Corresponding author at: Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, 501 Jack Stephens Drive, Mail Slot # 782, Little Rock, AR 72205, USA.
E-mail addresses: maryam.garza@duke.edu (M. Garza), guilherme.delfiol@utah.edu (G. Del Fiol), jessie.tenenbaum@duke.edu (J. Tenenbaum), acwalden@uams.edu (A. Walden), mzozus@uams.edu (M.N. Zozus).

http://dx.doi.org/10.1016/j.jbi.2016.10.016
1532-0464/

ity to satisfy stated or implied needs" [13]. CDM quality is conceptualized as the quality of a product (the model itself) rather than that of the process through which the model was created. With the exception of recent health informatics literature on the topic [5,14–17], data model evaluation research has focused almost solely on the use of data models as specifications.

Evaluation of a CDM's quality in terms of "fitness for use" can only be accomplished by comparing the model against the needs of the intended use. Our focus was on the utility of CDMs to support sharing of healthcare data, specifically longitudinal healthcare data for use in research.

1.3. Need for data model evaluation in clinical research

Multiple CDMs have been developed to support secondary use of healthcare data, including claims data in research; these include the Health Care Systems Research Network Virtual Data Warehouse, the OMOP CDM, Sentinel, and the PCORnet CDM. Multiple clinical research networks. Several registries have also developed data element sets or data models to facilitate their work [18,19]. CDMs have also been created to support healthcare uses of data such as clinical decision support (e.g., the Health Level 7 Virtual Medical Record) [20] and to facilitate alignment between multiple agencies (e.g., the Federal Health Information Model). However no CDMs have been evaluated or created for use in electronic health record (EHR)-based longitudinal registries. A confluence of two trends makes this an important problem to solve: (1) the specialization of CDMs, such as for clinical research versus healthcare and for claims versus EHR data, and (2) the increased exploration of observational, quasi-experimental, and experimental clinical study designs all using healthcare data.

2. Objective

We had a twofold objective: (1) to develop a generalizable methodology for the evaluation of CDMs, and (2) to apply the proposed methodology to a set of widely used CDMs in the context of a longitudinal EHR-based study.

3. Materials and methods

3.1. Evaluation process

The evaluation process entailed identification of candidate data models, configuring project-specific criteria, populating the candidate data model with test data, and applying the evaluation criteria (Fig. 1). The 300 EHR data elements for the MURDOCK (Measurement to Understand the Reclassification of Disease Of Cabarrus/Kannapolis) Community Registry and Biorepository were mapped to each candidate model, and each model was populated with test data. Mapping was independently verified for each model. Test data were used to evaluate each of the three queriability criteria for each model. Finally, the remainder of the evaluation criteria were applied to each model.

3.2. Support for longitudinal EHR-based registries

The MURDOCK Registry is a longitudinal, community-based health study working to enroll 50,000 consented adult residents in Kannapolis and Cabarrus County, NC, and the surrounding region (20 zip codes in total) [21]. The MURDOCK Registry is best described as a prospective cohort study. There are more than 12,000 participants actively enrolled in the registry at the time of this writing. All participants provide self-reported data, biosamples, permission to be re-contacted for future approved studies, yearly follow-up, and consent for longitudinal access to their health records. The registry captures data on (1) 34 self-reported medical conditions, hospitalizations, eight procedures, and medications in addition to demographic data and socioeconomic indicators, (2) corresponding health record data from multiple regional healthcare facilities, and (3) fully identified linkage of EHR data with self-reported data. The ability to use both participant self-reported data and data derived from clinical care in the participant's health record is a key aspect of the study design. This approach requires fully identified datasets to enable linkage of EHR data and self-reported data. The registry data are used for correlative studies with biospecimens, hypothesis generation, cohort identification, and eligibility screening for future studies. To be leveraged optimally for these purposes, the health record data must be shared in a structure that facilitates secondary use such as querying and report generation.

As such, the requirements for supporting longitudinal health record-based registries are:

1. Support fully identified health record data;
2. Support historical contact information;
3. Support regular follow-ups and the temporal aspect of longitudinal data;
4. Support common health record data domains such as demographics, social history, allergies and immunizations, diagnoses, procedures, problem list, encounters, hospitalizations, medications, labs, vital signs, and relationships as defined by the registry data model;
5. Support multiple sources of similar and related information, such as problems, diagnoses, and claims;
6. Support data linkage across primary, specialty, and hospital care settings;
7. Support both healthcare and research controlled terminologies;
8. Support operational requirements such as ease of querying, ease of implementation, low cost, and stability of the model in terms of user base, maintenance, and frequency of updates.

3.3. Selection of data models

Current versions of four candidate CDMs in use in clinical research were chosen for evaluation: Sentinel version 5.0, PCORnet version 3.0, OMOP version 5.0, and CDISC SDTM version 1.4 (Implementation Guide version 3.2). Our study revolved around the use of EHR data for clinical research, thus we limited the candidate data models to those designed and used for clinical research or their most current progeny. Therefore, data models designed to

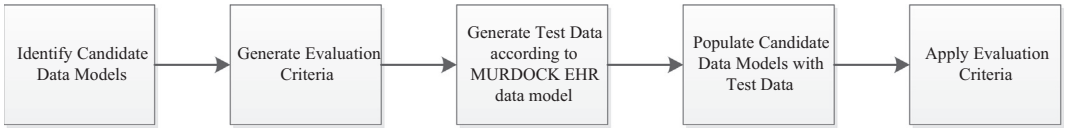

Fig. 1. Evaluation process.

support the provision of healthcare such as the Health Level Seven Reference Information Model, the Health Level 7 Virtual Medical Record, Open EHR, and the Federal Health Information Model, were not considered.

3.4. Evaluated models

Candidate models included those currently in use for clinical study data. Sentinel was developed as part of the Sentinel Initiative in 2008, in response to the Food and Drug Administration's (FDA's) efforts to create a national electronic system for monitoring the safety of FDA-regulated medical products [3]. As of January 2014, there are 22 Collaborating Institutions. Since its inception in 2008, there have been ten updates to the model, four of which were major version changes.

The open-source PCORnet CDM is based on the Sentinel CDM. PCORnet is a research partnership comprising clinical data research networks (CDRNs) and patient-powered research networks (PPRNs) [4]. There are currently 33 networks participating (13 CDRNs and 20 PPRNs) [4]. The first version of the CDM was released in 2014, and there have been two major releases and one minor update since then.

The OMOP is a public-private partnership "established to inform the appropriate use of observational healthcare databases for studying the effects of medical products" [6]. The OMOP community is composed of members from industry, government, and academia who actively use the OMOP CDM and vocabulary for their various research purposes [6]. The purpose of the OMOP CDM is to "standardize the format and content of observational data, so standardized applications, tools and methods can be applied to them" [6]. Version 1 was released as part of a pilot project in 2008, and since then, there have been 4 major releases.

CDISC SDTM was first released in 2004. It was selected as the standard specification for submitting tabulation data to regulating bodies, such as the FDA, for clinical trials (and for nonclinical studies since 2011) [7,8]. Version 1.0 was "designated as the first implementation-ready version for clinical studies involving human drug products" [7]. Subsequent versions are backward compatible, offering improvements and enhancements "to support a broader range of regulatory products" [7]. There are currently over 348 CDISC members.

3.5. Development of evaluation criteria

We developed and prioritized 11 criteria that were operationalized to EHR-based registries and demonstrated in our evaluation. The 11 evaluation criteria fell into six categories: content coverage (completeness), integrity, flexibility, ease of querying (simplicity), standards compatibility (integration), and ease and extent of implementation (implementability). Table 1 lists all criteria from the two sources alongside the generalized criteria for traceability.

Our development followed principles for structuring conceptual model quality frameworks based on the ISO/IEC 9126 standard. Briefly, the principles maintain that (1) conceptual model quality should be decomposed into a hierarchy of quality characteristics, subcharacteristics, and metrics, (2) single-word labels should be used for each quality characteristic and subcharacteristic, using commonly understood terms, (3) each quality characteristic and subcharacteristic should be defined using a single, concise sentence, (4) metrics should be defined for measuring each terminal characteristic, and (5) detailed procedures should be defined for conducting the evaluations.

The 11 evaluation criteria were identified from relevant published criteria and adapted by separating conceptual and operational definitions. The operational definitions were designed to be configurable—that is, generalized so that they can be applied to different uses. To develop the evaluation criteria, we synthesized two sets of criteria used and reported by Kahn et al. for their evaluation of data models for comparative effectiveness research (CER) [16]. The criteria lists from Kahn et al., which referenced the metrics developed by Moody and Shanks [22], were reviewed and matched to additional requirements to support longitudinal EHR-based registries. The developing requirement set was reviewed against the general questions enumerated by Ogunyemi et al. to assure that no additional areas of potential evaluation were missed [15]. Where necessary, additional criteria were developed to support EHR-based registries.

Conceptual definitions were harmonized across the sources for criteria. Subcriteria were added where two criteria were similar at some higher level of abstraction but conceptually different. For example, the completeness criterion was split into domain completeness and data element completeness based on our requirement for specific data elements. The data element completeness criterion was operationalized as percentage of data elements that mapped to a unique location in the evaluated data model. Operational definitions were adapted or drafted for each criterion.

Several criteria from the two sources were not carried forward, including correctness, scalability, understandability, and grid friendliness. Moody and Shanks define correctness as "whether the model conforms to the rules and best practices of the data modeling technique (i.e. whether it is a well-designed data model)" [22]. We did not assess whether a candidate model conformed to the underlying representational formalism as we did not find it critical to the evaluation compared to other criteria selected. The origin of the correctness parameter is the general Moody and Shanks criteria that stem from a desire to improve modeling in support of software development. In this context, model correctness is defined as the conformance of the model to the representational formalism; for example, does a UML model conform to the UML standard. All of the evaluated models are collaboratively

Table 1
Sources and generalization of evaluation criteria.

Evaluation criteria	Moody and Shanks[22]	Kahn et al.[16]	Generalized criteria
Completeness (Content Coverage)	✓	✓	✓
[1]Domains		✓	✓
[2]Data elements			✓
[3]**Integrity** (constraints & business rules)	✓		✓
Flexibility	✓	✓	✓
[4]Extensibility		✓	✓
Scalability		✓	
Adaptability		✓	
Understandability	✓	✓	
Correctness	✓	✓	
Simplicity (Ease of Querying)	✓		✓
Ease of transforming from storage to model views			
[5]Ease of querying the model views			✓
[6]Ease of anonymization and de-identification			✓
[7]**Integration (Standards Compatibility)**	✓	✓	✓
Implementability (Ease & Extent of Implementation)	✓	✓	✓
[8]Field experience		✓	✓
[9]Stability		✓	✓
[10]Adoption		✓	✓
Grid friendliness		✓	
[11]Cost		✓	✓

Note: The superscript numbers (1–11) indicate the 11 Evaluation Criteria chosen for this review. The 11 criteria fell into six categories, indicated by the bolded text: *completeness, integrity, flexibility, simplicity, integration, and implementability*.

developed and publically shared and implemented, therefore, some degree of technical correctness was assumed. Instead, we prioritized completeness of domain coverage and the availability of associations in a candidate model (integrity) that are necessary for EHR-based registries, as the effort required to modify or adapt a model for use would vary depending on the content currently captured. Similarly, scalability of the model itself was accounted for in the operationalization of ease of querying; essentially, run-time scales with number of joins and transactions interacting with the data. We retained complexity measures based on these rather than also retaining scalability as a criterion. While understandability becomes a factor with some very abstract models, this seemed less relevant to the models used in research, especially with the increase of the informatics workforce trained in data, information, and knowledge representation. Likewise, grid friendliness, while applicable to some uses, was not applicable to our specific use and thus not carried forward. Any of these, however, could be added to the generalized set to support specific project needs. Importantly, although additional subcriteria were developed, no new criteria were added to support our study, only more detailed operational definitions specific to evaluation of candidate data models for use with a given data set.

3.6. Conceptual and operational definitions of the criteria

The Kahn et al. and Moody and Shanks conceptual definitions of completeness were consistent. Like Kahn et al., seeking more specificity, we operationalized the completeness criteria to refer to content coverage of the data specified for EHR-based registries. Moody and Shanks defined integrity as the extent to which the data model "conforms to the business rules and processes to guarantee data integrity and enforce policies" [22]. Kahn et al. applied this definition to accommodate CER and defined integrity as the extent to which the data model enforces "meaningful data relationships and constraints that uphold the intent of the data's original purpose" [16]. We did the same.

Moody and Shanks defined flexibility as the extent to which the data model deals with changes in business or regulatory change [22]. Kahn et al., similarly but more specifically, defined flexibility as the extent to which new data elements and relationships can be added [16]. We further specified the addition of whole new domains. We did not include additions of new relationships because constraints (e.g., foreign key relationships) can generally be added without altering existing referential integrity, and presence of important associations was covered in the integrity criteria. We did not use Kahn et al.'s adaptability or scalability criteria.

Moody and Shanks defined simplicity as the extent the data model "contains the minimum possible entities and relationships" [22]. Kahn et al. expanded further by posing the following questions [16]:

1. Are concepts represented as straightforwardly as possible?
2. Are all data elements necessary?
3. Are the data elements easy to transform from the [electronic medical record] into the research data model?
4. Are the data elements easy to extract into a research dataset?

We constrained simplicity and defined it based on queriability, i.e., complexity of querying the data model. To determine the ease of querying, each candidate model was populated with test data. Several representative queries were developed and implemented in SQL against the candidate model. An example of one of our evaluation queries was "Number of registry participants listed as having received Narcan in an Emergency Department setting under 21 versus 21 and older."

We evaluated ease of de-identification and anonymization more subjectively. Each model was rated according to the amount of programming work required for our institutional anonymization expert to de-identify and anonymize the data. Essentially, the more complex the data model (i.e., the more data elements and domains), the more difficult it is to anonymize and de-identify the data. Time for reading and redaction of narrative fields to guard against indirect re-identification was not considered.

Moody and Shanks defined integration as the extent to which the model is consistent with the rest of the organization's data [22]. Kahn et al. further specified the definition [16]. For our study, the evaluation focused primarily on the use of standardized vocabularies with which we had already aligned: RxNorm, NDF-RT (National Drug File Reference Terminology), ICD (International Classification of Diseases), CPT (Current Procedural Terminology), HCPCS (Healthcare Common Procedure Coding System), SNOMED-CT (Systematized Nomenclature of Medicine--Clinical Terms), LOINC (Logical Observation Identifier Names and Codes), NDC (National Drug Code), and NPI (National Provider Identifier).

Both Moody and Shanks and Kahn et al. define implementability as the ability of the model to be implemented (and maintained) within the anticipated time, budget, and technical constraints [16,22]. Kahn et al. further specify implementability in terms of five subcriteria: field experience, stability, adoption, grid friendliness, and cost [16]. We retained field experience, stability, adoption, and cost as criteria. The resulting conceptual and operational definitions are presented in Table 2.

4. Results

4.1. Completeness – content coverage

Of the four candidate models, OMOP had the highest domain coverage (76% of the data elements), followed by SDTM at 55%, PCORnet at 48%, and Sentinel at 37% (Table 3). While SDTM accommodated a significant percentage of data elements (55% true matches), 45% of the data elements did not match a topicality-based SDTM domain and thus were mapped to the Supplemental Qualifiers (SUPPQUAL) domain, an extension mechanism utilized for representing and relating data values not accommodated in an existing domain [7]. Because SUPPQUAL is a highly normalized Entity-Attribute-Value-like (EAV) domain, it requires additional joins at the time of data use to rejoin the data elements with other data of common topicality. Thus, SDTM extensibility significantly increases burden at the time of data use, while data elements not mapping to OMOP core tables do not. For example, the OMOP extension strategy allows addition of data elements to existing domains rather than mapping them to a "catch-all" EAV table. SDTM relegates the non-mapping data elements to SUPPQUAL. Data elements mapped to SUPPQUAL were not counted as accommodated in our overall evaluation. The results are presented by category so that readers can combine in ways applicable to other projects.

4.2. Integrity – model associations

Both OMOP and SDTM had a 100% match of the data model constraints, whereas Sentinel matched 38% and PCORnet matched 66% (Table 4). While SDTM matched all of the MURDOCK associations, 86% of those associations incorporated the SUPPQUAL domain.

4.3. Flexibility – extensibility

All of the candidate data models rated similarly in flexibility and extensibility (Table 5), as all three models are able to

Table 2
Conceptual and operational definitions of the 11 evaluation criteria.

Criterion	Conceptual definition	Operational definition
Completeness – [1]*Data elements*	All data elements for the projects are accommodated by the model	(1) Percentage of data elements that map to core domains or tables in the evaluated model, (2) Percentage of data elements that do not map to any tables in the evaluated model, and (3) The percentage of data elements that map to non-core domains, tables, or model extensions in the evaluated model
Completeness – [2]*Domains*	All domains are accommodated by the model	Percentage of domains required for the project that are captured by the evaluated model
[3]**Integrity**	The extent to which associations in the evaluated data model match the specific project needs	(1) The percentage of associations in our data model not held by the evaluated model, (2) The number of additional constraints specified in the evaluated data model not required by the project (not evaluated because we were comfortable with relaxing these constraints), and (3) Whether the evaluated model accommodated a data change log and data provenance information
[4]**Flexibility**	The ease with which the evaluated model can adapt to changes such as adding new data elements and domains without changing referential integrity	(1) Whether a new data element can be added without changing referential integrity (i.e., adding an attribute to an entity is acceptable), and (2) Whether a new domain can be added without changing referential integrity as described in the extensibility definition above
Simplicity – [5]*Queryability*	The ease of querying the evaluated model for cohort identification	(1) Number of table joins required for each query, (2) Number of nested queries needed for each query, and (3) Qualitative estimate (faster, or slower) of the overall query performance over views of the data model based on the complexity of the query used for cohort identification These assume a relational database implementation
Simplicity – [6]*De-identification & Anonymization*	The ease of de-identification and anonymization of the data captured in the evaluated model	Qualitative estimate (easy, medium, or difficult) of the complexity of the de-identification and anonymization process
[7]**Integration**	The extent to which the model supports controlled terminologies	Each model was evaluated on the integration and use of a controlled vocabulary. The terminologies supported by the models were compared to the seven different vocabularies we use. The instances where the model precluded us from using the controlled terminology were documented, and each model was given a percentage based on the number of terminologies that matched
Implementability – [8]*Field Experience*	The number and diversity of uses of the data model [16]	(1) Diversity of intended use based on the model's original intent, and (2) The number of years the model has been in use
Implementability – [9]*Stability*	The total number of changes to the data model	Count of model updates in the last 2 years
Implementability – [10]*Adoption*	The size of the user community using and supporting the data model	(1) The total number of adopters for each model, and (2) The percentage of MURDOCK data users that use that particular model
Implementability – [11]*Cost*	The licensing fees, staffing, and costs of infrastructure to implement, maintain, and use the data model	This criteria was broken down into four components, each evaluated independently: (1) licensing fees for the model itself and any required controlled terminology, (2) programming time estimate, (3) hardware and software cost, and (4) training time Criteria 2–4 were assessed by software engineers in the MURDOCK team according to a 3-point scale: *little*, *medium*, or *a lot*

Note: The superscript numbers (1–11) indicate the 11 Evaluation Criteria chosen for this review. The 11 criteria fell into six categories, indicated by the bolded text: *completeness, integrity, flexibility, simplicity, integration, and implementability*.

Table 3
Content coverage results.

Evaluation criteria	Sentinel	PCORnet	OMOP	CDISC SDTM
% of data elements that map to core tables	37%	48%	76%	55%
% of data elements do not map to core tables	63%	52%	24%	n/a
% of data elements that map to model extensions	n/a	n/a	n/a	45%

Table 4
Integrity results.

MURDOCK associations	Sentinel	PCORnet	OMOP	CDISC SDTM
% of associations that match MURDOCK	38%	66%	100%	100% (86%)[a]
Demographics – Contact Info	NO	NO	YES	YES[a]
Demographics – Social History	NO	NO	YES	YES[a]
Demographics – Encounter	YES	YES	YES	YES[a]
Demographics – Patient Medical History	NO	YES	YES	YES[a]
Demographics – Allergies	NO	NO	YES	YES[a]
Demographics – Immunization History	YES	NO	YES	YES[a]
Demographics – Labs	YES	YES	YES	YES[a]
Demographics – Microbiology	YES	YES	YES	YES[a]
Demographics – Diagnosis	YES	NO[b]	YES	YES[a]
Demographics – Procedures	YES	YES	YES	YES[a]
Demographics – Problem List	NO	YES	YES	YES[a]
Demographics – Medications	NO	YES	YES	YES[a]
Demographics – Vitals	YES	YES	YES	YES
Encounter – Social History	NO	NO	YES	YES[a]
Encounter – Hospitalization	YES	YES	YES	YES
Encounter – Vitals	NO[c]	YES	YES	YES
Encounter – Medications	NO[c]	YES	YES	YES[a]
Encounter – Labs	NO[c]	YES	YES	YES[a]
Encounter – Microbiology	NO[c]	YES	YES	YES[a]
Encounter – Diagnosis	YES	YES	YES	YES[a]
Encounter – Procedures	YES	YES	YES	YES[a]
Encounter – Problem List	NO	YES	YES	YES[a]
Encounter – Allergies	NO	NO	YES	YES[a]
Encounter – Immunization History	NO[c]	NO	YES	YES[a]
Encounter – Patient Medical History	NO	YES	YES	YES[a]
Patient Medical History – Family Medical History	NO	NO	YES	YES[a]
Labs – Microbiology	YES	YES	YES	YES[a]
Microbiology – Microbiology Results	NO	NO	YES	YES[a]
Microbiology Results – Microbiology Sensitivity	NO	NO	YES	YES[a]

[a] Associations that incorporate the SUPPQUAL domain.
[b] Medications in Sentinel are captured in the Outpatient Pharmacy Dispensing table in which "each record represents an outpatient pharmacy dispensing" [3]. Prescription data are not currently captured in the Sentinel CDM, and, as outpatient dispensing data is not commonly captured within healthcare systems, it was not considered a match in our project.
[c] Date Windowing: Indicates that while there was no direct association, approximate (non-exact) associations could made by matching encounter date with the observation date.

Table 5
Flexibility, simplicity, integration, and implementability results.

Evaluation criteria	Sentinel	PCORnet	OMOP	CDISC SDTM
Flexibility – extensibility (ease of adding new data elements)	YES	YES	YES	YES
Simplicity				
No. of nested queries	1	1	1	1
No. of table joins	1–5	1–6	2–5	4–12
Estimated query performance	Faster	Faster	Faster	Slower
Complexity of anonymization / de-identification	Easy	Medium	Medium	Difficult
Integration				
% same controlled terminology used	67%	78%	100%	67%
No. of controlled terminology the model does not support	3	2	0	3
Implementability – field experience (no. of years the model has been in use)	8	2	7	12
Implementability – stability				
Model updates in the last 2 years	1	3	1	0
Minor clarifications/text modifications	0	1	0	0
Major changes	1	2	1	0
Implementability – adoption				
No. of adopters	18[a]	33[b]	5[c]	>348
% of MURDOCK data exchange users	85%	85%	85%	13%
Implementability – cost				
Licensing fees	$0	$0	$0	$0
Programming time estimate	Some	Some	Little	A lot
Hardware and software costs	None	None	None	None
Training time	Little	Little	Some	Some

[a] If members within larger organizations, such as Kaiser and the Health Care Systems Research Network (formerly HMO Research Network), are counted separately, the count would be 31 adopters.
[b] There are 33 partner networks (13 CDRNs and 20 PPRNs) comprised of various healthcare systems, providers, and patients[4].
[c] Five adopters were found using Internet searches; however, no list was available. Thus, this count should be interpreted as a lower bound.

accommodate adding new data elements. However, PCORnet, Sentinel and OMOP are better suited to adding new domains. SDTM's strategy for adding new domains is based on template findings, interventions, and events domains, and relies on SUPPQUAL for data that do not map to the templates.

4.4. Simplicity – ease of querying

Sentinel, PCORnet, and OMOP are equivalent with regard to the number of nested queries, table joins, and overall query performance for the evaluated queries. SDTM performed slightly less favorably, primarily because many queries would require joining to the SUPPQUAL table.

4.5. Integration – standards compatibility & terminology coverage

OMOP had the best terminology coverage at 100%, matching all nine terminologies required by our project described above in the Methods section. PCORnet covered 78% with seven out of nine matches, while Sentinel and SDTM covered 67%, with six out of nine matches.

4.6. Implementability – field experience, stability, adoption, and cost

Of the four models, the SDTM CDM has been available the longest (12 years), followed by Sentinel (8 years), OMOP (7 years), and PCORnet (2 years). Three of the four models have seen revisions in the past 2 years. PCORnet has had the most revisions (two major changes and one minor update), while OMOP and Sentinel have each had one major change (Table 5). SDTM has the highest number of adopters (>348), but Sentinel, PCORnet, and OMOP have a greater percentage of MURDOCK data exchange users (academics vs industry collaborators). Past, current, and prospective MURDOCK Study collaborators are primarily academic and therefore more likely to use Sentinel, PCORnet, or OMOP than SDTM.

In general, all four models evaluated were rated similarly with regard to cost, but with important differences in programming and training time. Specifically, OMOP was rated more favorably in programming time estimate, since it captures most of the data elements and domains. It is important to note that we did not estimate costs for mapping to standard vocabularies. While we recognize that vocabulary mapping typically requires substantial effort, this would not be influenced by the individual data model. This effort is required regardless of the target model and will typically occur outside the model (i.e., prior to the ETL). Therefore, cost estimates of this effort were considered to be outside the scope of this paper.

5. Discussion

Multiple CDMs exist for clinical research and EHR data. However, no CDMs have been evaluated or developed for use in EHR-based longitudinal registries such as the MURDOCK Community Registry and Biorepository. At a minimum, efficient evaluation methodology is needed so that available CDMs can be systematically identified and evaluated for projects across the spectrum of clinical research and other secondary data uses. Evaluation methodology will increase availability of actionable feedback for model improvement and possible convergence or development of transformations between the models so that, where possible, data shared in different models can be computationally pooled in fully or in semi-automated fashion, ultimately decreasing the cost of data sharing and re-use. In the present study, we address this gap by proposing and applying a methodology to evaluate CDMs in the context of EHR-based longitudinal registries. Although our study was driven by the requirements of the MURDOCK Registry, the methodology is generalizable to other studies.

Putting the methodology presented here in the context of related work, like Kahn et al., our work used select Moody and Shanks metrics. Like Overhage et al., our evaluation leveraged realistic clinical data and mapped the data into candidate models, using the results to calculate the metrics [14]. Like Recker and Rosemann, we heavily relied on Bunge-Wand-Weber (BWW) categories (categorizing the metrics according to BWW classes) [23]. However, diverging from Recker and Rosemann, we operationalized the metrics for direct quantification for a majority of the criteria. For example, our assessment of completeness tallied the number of data elements from our model with a one-to-one mapping to each candidate data model. Following BWW assured that our criteria and resulting operationalizations (metrics) did not conceptually overlap and that the set was complete. Like Moody and Shanks and Kahn et al., we added additional criteria because, whereas BWW is focused on representational correctness of a domain, we had additional considerations such as implementation cost and likelihood of continued support for the model.

Ultimately, the BWW representational theory criteria stood out as a minimum that a candidate CDM needed to meet to warrant further consideration. Representation expressivity has been defined by Wand and Weber [24] as the ability to describe real-world phenomena completely and clearly (operationalized respectively as content coverage including contextual associations, and one-to-one mapping of singular concept data elements from the source data to the target CDM). A model that lacks representation expressivity degrades the data through reduction or altogether exclusion. Thus, in practice, evaluation of CDMs for a particular use is a two-stage process where the first stage identifies models with adequate content coverage and associations, and the second stage assesses other considerations of importance to the specific project or study.

Our work adds to previous related efforts on integrated data repositories (IDRs). MacKenzie et al. [25] evaluated commonalities in IDR implementation and use among several major academic health centers using IDRs to support clinical research. The survey collected information on the following areas: (1) general IDR status, (2) data sources supporting the IDR, (3) technical architecture and staffing, (4) clinical systems, and (5) project experience [25]. Comparing survey results between 2008 and 2010, the study found a shift towards centralized IDRs; a shift away from home-grown systems to "more commonly used platforms"; IDRs being used to store core clinical data elements, rather than solely administrative data; and data from multiple institutions being stored within IDRs. While MacKenzie et al. focused on identifying the common challenges faced by institutions implementing and using IDRs for clinical research, our study focused on the development and application of criteria to evaluate CDMs for research.

Huser and Cimino also approached the much related problem of restructuring repositories for data from healthcare settings. An architectural comparison was performed of three large IDRs: i2b2, HMO Research Network Virtual Data Warehouse (VDW), and OMOP. Specifically, Huser and Cimino focused their evaluation on two aspects: (1) architecture for storing facts, and (2) structures for representing the terminology layer of the warehouse [26]. Their analysis resulted in the formulation of a set of desirable characteristics ("desiderata") that have general applications across IDRs [26]. Several methodological differences exist between our investigation and that of Huser and Cimino. Mainly, while our methodology is generalizable for the evaluation of CDMs, we focused our evaluation on data models for EHR-based longitudinal registries. Huser and Cimino did not approach their investigation with any particular use in mind, but focused rather on storage and organization of healthcare data to support clinical research in general.

While some may use CDMs to structure IDRs, this may not always be the case. In fact, in our work, we assumed the CDM would be for sharing data with others and not necessarily for the structure of the data storage. However, we acknowledge that some may implement CDMs as the structure for an IDR. Therefore, the two approaches are complementary and can be used in tandem to assess the suitability of specific CDMs or IDRs for specific applications.

Potential limitations of CDMs for data originating from other systems include unmappable data, information loss associated with abstracting above individual source system differences, and differences in supported associations. As articulated by Kahn et al., the impact of data modeling decisions on use of data in clinical research is not well documented [16]. Modeling decisions affect not only the amount of data in a specific area that are covered, but also how the data are related. For example, diphenhydramine may be prescribed for a suspected immune reaction to a drug. Some data models may preserve this relationship (drug A administered for suspected immune reaction to drug B), while others may not. Data model evaluation should account for necessary data, their representation, and their relationships, and in this way can guide model selection for a particular study. As pointed out by Overhage et al., mapping data to a CDM may preclude maintaining some of the relationships or data contained in the original data [14].

At one point in this work and in previous work [27], we considered use of CDMs as a logical data model for data storage. In both cases, we found the transformations to the CDMs to be "lossy" in terms of reduction of the information content of the data. Marital status mapping from source EHR data to the OMOP models provides a simple example. One of the source EHRs from which our registry obtained data had a single category for married, while the OMOP model uses three categories—married, legally married, married/civil partner. Similarly, the OMOP model contained several categories for a single person, including single person and single, never married. These mapping mismatches occurred in both directions—that is, (i) source data were more detailed, requiring multiple values to be mapped to a single CDM category, and (ii) source data were less detailed, making it impossible to assign the values to the finer-grained CDM categories. Similar bidirectional challenges occurred with associations: some associations in source data were not available in the CDM, and in other cases, the source data did not contain associations that were enforced in the CDM. We did not encounter discretization of continuous data in the work here but have encountered discretization-associated data reduction in practice. Unless a CDM is a perfect representation of source data, information loss will result. For these reasons, we do not recommend use of a CDM for data storage unless the CDM was developed for that particular use at hand or has evaluated extremely for content coverage. Instead, we favor of less "lossy" approaches such as those reported by Cimino et al. for data storage [28]. We emphasize here the use of data models for sharing data and combining data from multiple sources rather than for data storage. Models that are project specific inevitably result in data reduction, which can preclude future unplanned uses of the data and ultimately decrease the data's usefulness. Thus, mapping to a CDM should occur at the latest practicable stage closest to the analysis to preserve the full information content of the data for potential secondary uses.

There are limitations to the evaluation methodology proposed here. First and foremost, CDMs in clinical research were developed for specific uses, some more generalized than others. A CDM designed for one use might not support a different but seemingly similar use; therefore, CDM evaluation methodology necessarily must facilitate project-specific criteria. Our methodology has not been tested with other uses, and thus generalizability remains to be confirmed. In keeping with earlier work, some of the criteria are subjective. Rating for those criteria was done only by a few members of the MURDOCK team and in practice could be strengthened by operationalizing the criteria more objectively or by having multiple individuals assess the subjective criteria. Lastly, the criteria and their synthesis from the literature have not been externally validated. Further refinement by an expert panel could be beneficial.

6. Conclusion

Existing CDMs were developed for specific uses, and their fitness for other uses depends on how closely the CDM matches the planned use. Thus, methodology for project-specific evaluation of CDMs is needed. To this end, we synthesized a set of 11 criteria for evaluation of CDMs. The 11 criteria fell into six categories: content coverage (completeness), integrity, flexibility, ease of querying (simplicity), standards compatibility (integration), and ease and extent of implementation (implementability). The criteria and methodology reported here support evaluation in specific contexts through the ability to ignore criteria unimportant to a specific use and add criteria where needed.

We tested the methodology to evaluate CDMs for EHR-based longitudinal registries. Overall, OMOP performed best in the evaluation, ranking highest in a majority of the evaluation criteria compared with the other models – content coverage, integrity, flexibility, simplicity, integration, and implementability. The OMOP model accommodates the highest number of our data elements and has the broadest coverage of standard terminologies. Based on our evaluation, the OMOP CDM will require less configuration, training, and modifications to support EHR-based registries.

Based on the observed data reduction in mapping to the various CDMs, we conclude that transformation of data to CDMs should occur as late in the data processing stream as possible, that is, as close to the analysis as possible. We favor least lossy approaches for data storage and management.

Contributors

M.Y.G.: Contributed to methodology design; Conducted evaluation and analysis, wrote the manuscript.

G.D.F., J.D.T., A.W.: Contributed to methodology design and the analysis and interpretation of the data; critically reviewed and revised manuscript.

M.N.Z.: Designed methodology; contributed to analysis and interpretation of the data; wrote the manuscript.

Funding

This work was supported by a gift to Duke University from the David H. Murdock Institute for Business and Culture and Duke University's CTSA grant (UL1TR001117) from the National Institutes of Health (NIH) National Center for Advancing Translational Sciences, grant UG1DA040317 from the NIH National Institute on Drug Abuse, and R00-LM011128 from the NIH National Library of Medicine. The ideas presented here do not necessarily represent the David H. Murdock Institute for Business and Culture or the NIH, and are the sole responsibility of the authors.

Competing interests

None.

Acknowledgments

The authors would like to thank the following individuals and groups for their contributions: MURDOCK Study Technical Team: Julie Frund, Anne Heath, Kurt Morehouse.

Data Model Mapping Reviewers: Lesley Curtis (Sentinel), Rhonda Facile (CDISC SDTM), Amy Palmer (CDISC SDTM), Shelley Rusincovitch (PCORnet), Yinghong Zhang (Sentinel and OMOP).

Editor: Peter Hoffmann, Duke Clinical Research Institute.

References

[1] M. Mosley (Ed.), The DAMA Dictionary of Data Management, first ed., Technics Publications, Denville, NJ, 2008.
[2] M. Abdelhak, S. Grostick, M.A. Hanken, Health Information: Management of a Strategic Resource, third ed., Elsevier Health Sciences, St. Louis, MO, 2007.
[3] Sentinel Common Data Model v5.0. <http://www.mini-sentinel.org/> (accessed 1 February, 2016).
[4] The National Patient-Centered Clinical Research Network (PCORnet) Common Data Model v3.0. <http://www.pcornet.org/resource-center/pcornet-common-data-model/> (accessed 1 February, 2016).
[5] T.R. Ross, D. Ng, J.S. Brown, et al., The HMO research network virtual data warehouse: a public data model to support collaboration, EGEMS 2 (1) (2014) 1049. Doi: 10.1306372327-9214.1049. eCollection 2014.
[6] Observational Medical Outcomes Partnership (OMOP) Common Data Model v5.0. <http://omop.org/> (accessed 14 October, 2014).
[7] Clinical Data Interchange Standards Consortium, Submission Data Standards Team, Study Data Tabulation Model (v1.4): 1–40. <http://www.cdisc.org/sdtm>, 2013 (accessed 1 June, 2014).
[8] Clinical Data Interchange Standards Consortium, Submission Data Standards Team, Study Data Tabulation Model Implementation Guide: Human Clinical Trials (v3.1.2): 1–398. <http://www.cdisc.org/sdtm>, 2013 (accessed 1 June, 2014).
[9] S. Olson, A.S. Downey (Eds.), Sharing Clinical Research Data: Workshop Summary, National Academies Press, Washington, DC, 2013.
[10] NIH Data Sharing Policy and Implementation Guidance. <http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm> (accessed 1 July, 2015).
[11] National Science Foundation, Grant Proposal Guide, NSF 11-1 January, 2011, Chapter II.C.2.j. <http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_index.jsp> (accessed 30 June, 2015).
[12] Organisation for Economic Co-operation and Development (OECD), Science, Technology and Innovation for the 21st Century, Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29–30 January, 2004 – Final Communique. <http://www.oecd.org/science/sci-tech/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm> (accessed 1 July, 2015).
[13] D.L. Moody, Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions, Data Knowl. Eng. 55 (2005) 243–276.
[14] J.M. Overhage, P.B. Ryan, C.G. Reich, et al., Validation of a common data model for active safety surveillance research, J. Am. Med. Inform. Assoc. 19 (1) (2012) 54–60.
[15] O.I. Ogunyemi, D. Meeker, H.E. Kim, et al., Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems, Med. Care 51 (8 Suppl 3) (2013) S45–S52.
[16] M.G. Kahn, D. Batson, L.M. Schilling, Data model considerations for clinical effectiveness researchers, Med. Care 50 (Suppl) (2012) S60–S67.
[17] Y. Xu, X. Zhou, B.T. Suehs, et al., A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: implications for active drug safety surveillance, Drug Saf. 38 (8) (2015) 749–765.
[18] HL7 Version 3 Domain Analysis Model: Emergency Medical Services, Release 1, HL7 Informative Document: HL7 V3 DAM EMS R1-2013. Available from <www.HL7.org: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=39> (accessed on 1 July, 2015).
[19] HL7 Version 3 Domain Analysis Model: Trauma Registry Data Submission, Release 1, July 2014, HL7 Informative Document: HL7 V3DAM TRAUMA, R1, A Technical Report prepared by Health Level Seven International and Registered with ANSI: 8/24/2014. Available from <www.HL7.org: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=363> (accessed on 1 July, 2015).
[20] K. Kawamoto, G. Del Fiol, H.R. Strasberg, et al., Multi-national, multi-institutional analysis of clinical decision support data needs to inform development of the HL7 virtual medical record standard, in: AMIA Annu Symp Proc. 2010 November 13, 2010, pp. 377–381.
[21] J.D. Tenenbaum, V. Christian, M.A. Cornish, et al., The MURDOCK study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records, Am. J. Transl. Res. 4 (3) (2012) 291–301.
[22] D.L. Moody, G.G. Shanks, Improving the quality of data models: empirical validation of a quality management framework, Inform. Syst. 28 (2003) 619–650.
[23] J. Recker, M. Rosemann, Measuring perceived representational deficiencies in conceptual modeling: instrument development and test, in: 29th International Conference on Information Systems (ICIS 2008): Proceedings of a Meeting Held 14–17 December 2008, Paris, France, 2008.
[24] Y. Wand, R. Weber, On the ontological expressiveness of information systems analysis and design grammars, Inform. Syst. J. 3 (4) (1993) 217–237.
[25] S.L. MacKenzie, M.C. Wyatt, R. Schuff, et al., Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey, J. Am. Med. Inform. Assoc. 19 (2012) e119–e124.
[26] V. Huser, J.J. Cimino, Desiderata for healthcare integrated data repositories based on architectural comparisons for informatics development of three public repositories, AMIA Annual Symposium Proceedings, vol. 2013, 2013, pp. 648–656.
[27] N. Hayes, P. Warfel, O. Oladapo, et al., CDISC Implementation Mostly in Clintrial™ at the Duke Clinical Research Institute, Society for Clinical Data Management Annual Meeting, Toronto, Ontario, Canada, September 2004.
[28] J.J. Cimino, E.J. Ayres, L. Remennik, et al., The national institutes of health's biomedical translational research information system (BTRIS): design, contents, functionality and experience to date, J. Biomed. Inform. 52 (2014) 11–27.