Data Model Considerations for Clinical Effectiveness Researchers

Michael G. Kahn, MD, PhD,†‡ Deborah Batson, BS,‡ and Lisa M. Schilling, MD, MSPH§*

Introduction: Growing adoption of electronic health records and increased emphasis on the reuse and integration of clinical care and administration data require a robust informatics infrastructure to inform health care effectiveness in real-world settings. The Scalable Architecture for Federated Translational Inquiries Network (SAFTI Net) was one of 3 projects receiving Agency for Healthcare Quality and Research funds to create a scalable, distributed network to support Comparative Effectiveness Research. SAFTINet's method of extracting and compiling data from disparate entities requires the use of a shared common data model.

Data Models: Focusing on the needs of CER investigators, in addition to other project considerations, we examined the suitability of several data models. Data modeling is the process of determining which data elements will be stored and how they will be stored, including their relationships and constraints. Addressing compromises between complexity and usability is critical to modeling decisions.

Case Study: The SAFTINet project provides the case study for describing data model evaluation. A sample use case defines a cohort of asthma subjects that illustrates the need to identify patients by age, diagnoses, and medication use while excluding those with diagnoses that may often be misdiagnosed as asthma.

Discussion: The SAFTINet team explored several data models against a set of technical and investigator requirements to select a data model that best fit its needs and was conducive to expansion with new research requirements. Although SAFTINet ultimately chose the Observation Medical Outcomes Partnership common data model, other valid options exist and prioritization of requirements is dependent upon many factors.

From the *Department of Pediatrics, Division of Pediatric Epidemiology; †Colorado Clinical and Translational Sciences Institute, University of Colorado, Denver; ‡Department of Clinical Informatics, Children's Hospital Colorado; and §Department of Medicine, Division of General Internal Medicine, University of Colorado, Denver, Aurora, CO.

Funding was provided by a contract from AcademyHealth. Additional funding was provided by AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network) and NIH/NCRR Colorado CTSI Grant Number UL1 RR025780 (Colorado Clinical and Translational Sciences Institute).

The authors declare no conflict of interest.

Reprints: Michael G. Kahn, MD, PhD, Children's Hospital Colorado, 13123 East 16th Avenue, B400, Aurora, CO 80045. E-mail: michael.kahn@ucdenver.edu.

ISSN: 0025-7079/12/5007-0S60

Key Words: data models, databases, Comparative Effectiveness Research

(*Med Care* 2012;50: S60–S67)

In the United States, the adoption of electronic health record systems is expected to increase rapidly, spurred by financial incentives and penalties mandated in the American Recovery and Reinvestment Act.[1–3] At the same time, interest has grown in evaluating the effectiveness of clinical care practices using electronic data recorded during routine clinical care.[4,5] Under the title of Comparative Effectiveness Research (CER), these studies focus on health outcomes, clinical effectiveness, risks, and benefits of medical care in real-world practice settings using clinical, administrative, and billing data.

As part of American Recovery and Reinvestment Act's $1.1B funding allocation, the Agency for Healthcare Quality and Research funded multiple projects focused on building scalable, distributed research networks to support CER.[6] The Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) received funding through this program. SAFTINet's goal is to build a distributed research network to support CER with a focus on safety-net stakeholders, which includes persons lacking health insurance and those with Medicaid and State Children's Health Insurance Programs. SAFTINet will combine detailed clinical and financial data from electronic health records, state Medicaid claims, and administrative data sources into a secondary analytic-only database that is separate from databases in clinical applications to answer questions regarding the comparative effectiveness of treatments, diagnostics, protocols, and other delivery system factors. Like many beginning CER projects, SAFTINet needed to either develop or adopt a data model for storage, processing, and analysis of a broad range of clinical and financial data. Using SAFTINet as a case study, we describe the considerations that were applied during our quest to identify an acceptable data model.

WHAT IS A DATA MODEL AND WHY IS IT IMPORTANT

Data modeling is the process of determining how data are to be stored in a database.[7–9] A data model specifies features and relationships, such as:

- Data types (eg, date, integer, character, time)
- Constraints (Are missing values allowed? Must each value be unique?)

- Relationships between rows of data (Can a row in 1 table be related to none, 1, or many rows in another table? Can hierarchies that define sets of concepts be represented?)
- Metadata definitions, procedures, and assumptions that describe the intended meaning and use of each data element, how data are to be collected, allowed values or ranges, and dependencies between data elements.

The structure and metadata definitions contained in a data model heavily influence what research data can be stored, how data values should be interpreted, and how easily desired data subsets can be queried and extracted from a research database.

Despite a large computer and information sciences-oriented technical literature on data modeling,[8,10–13] the choices, options, and impact of data modeling decisions to support *clinical research* are neither well studied nor published.[14,15]

The structural components of a data model typically are conveyed schematically in drawings that use symbols and notations to denote the features of and relationships among data items. Figure 1 illustrates 2 very simple data models drawn in a widely used format called the Entity-Relationship Diagram (ERD). The diagrams show each item's data type (integer, characters, dates), notes if the data element is required to always have a value (NN = never null, ie, it cannot be null/missing), and if it is a primary (PK) or foreign key (FK). For example, the patient table in model 1a has a data item named "PAT_ID" that is an integer, must always have a value (NN), and is a PK. ERDs capture the database structure but not metadata (descriptions of the data meanings) that

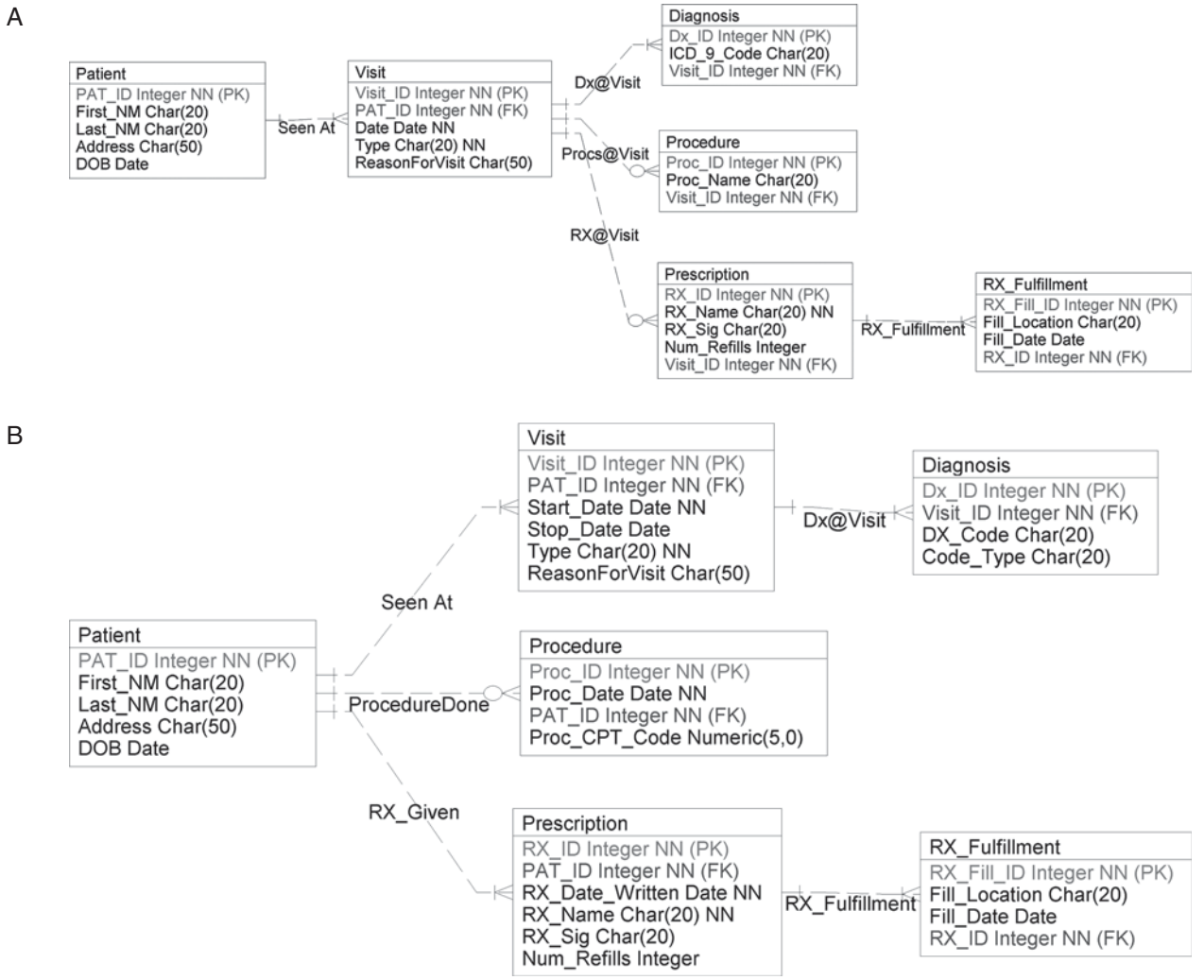

FIGURE 1. Two alternative simplified data models for representing clinical data. NN indicates never null.

usually is captured in related documents that are linked to an ERD.

Despite their simplicity, the ERDs in Figure 1 visually illustrate markedly different assumptions in how 2 models organize and relate data items. Model 1a is *visit-centric*, requiring all diagnoses, procedures, and medication prescriptions to be associated with a visit. None of these entities contain a date field because the date is obtained from the required associated visit. On the basis of the symbols at the ends of the connecting lines (lines, circles, and crow's feet), the model 1a "says" that a visit does not always need to have an associated procedure or a prescription but must have *at least one* ICD-9-CM diagnosis. In contrast, model 1b is *patient-centric*, associating visits, procedures, and prescriptions with a patient rather than a visit. In this model, procedures and prescriptions contain an independent date field. Data in this field can be unrelated to any visit, allowing for procedures and prescriptions to be recorded when there was no corresponding visit. The first model also has a single date field for visits whereas the second model has 2 date fields for visits. Model 1b allows for both hospital admission and discharge dates to be captured whereas model 1a would require a decision about which date to store in the single visit date field. Diagnoses in model 1a can only be ICD-9-CM codes whereas model 1b can store any type of diagnosis codes (ICD-9-CM, ICD-10-CM, SNOMED) using different values in Code_Type to distinguish between different coding systems. CPT codes can be recorded in model 1b but not in model 1a.

Multidisciplinary teams including informatics professionals, clinical investigators, and biostatisticians use ERDs to evaluate a proposed data model's ability to meet the data storage and querying needs of a research project. A significant challenge for CER projects is designing a database structure that is sufficiently flexible to support a wide variety of data types collected from electronic health records, billing systems, pharmacy dispensing, and pharmacy benefits/claims processing systems. The process of combining related data from disparate sources that have different data structures, variable formats, definitions, specificity, and quality, is called data integration.[16–19] Data integration requires careful attention to how the data from different systems will be represented in the research data model and how differences in data definitions, procedures, and sources will be resolved to ensure compatibility and comparability of the resulting data values. As data integration decisions and data model structures and definitions are so closely intertwined, these decisions are best addressed when investigators and analysts (those who will be using the database to answer questions) work closely with database designers (those who model and create the database) to explore the impacts of design decisions on database maintenance and extensibility, data quality, accessibility, and analytic capability.

The 2 simple data models in Figure 1 embody different assumptions. Model 1a assumes all actions that *matter to the study* occur during a visit. Model 1b is more general because it represents actions both during and between visits. However, for model 1a, the query: "What is the average number of prescriptions written per ambulatory visit?" is trivial because of the direct connection between visits and prescriptions. Answering the same question with model 1b is more complex, requiring a comparison of dates between visits and prescriptions to find prescriptions written on the same date as a visit.

Balancing flexibility versus complexity is a common tension in data modeling. A simple data model may not record important details; a more complex model may be more difficult to query. Given inherent tradeoffs, the next question to explore is: "How does one judge the acceptability or quality of a data model for a particular use?"

EVALUATING DATA MODEL QUALITY

Numerous approaches to assessing data models appear in the Information Systems literature.[20–26] In an examination of data model alternatives for the Food and Drug Administration Sentinel Initiative,[15] Brown and colleagues list 5 key related questions:

1. What does the system need to do?
2. What data are needed to meet system needs?
3. Where will the data be stored?
4. How will the data be analyzed?
5. Is a common data model needed, and if so, what will the model look like?

Although the 5 questions selected by Brown may be the most important questions *in their specific setting*, they may not to be the most important in other settings. Moody and Shanks created a comprehensive framework to ensure that all potential features of a data model are considered.[27–29] They proposed 8 dimensions to be considered in evaluating a data model.[28] We have restated the 8 dimensions into a CER context (Table 2). The Moody and Shanks framework emphasizes an *integrated* analysis—each setting will consider some factors to be critical, other factors to be "nice to have," and the remainder to be not relevant.

IDENTIFYING AND EVALUATING DATA MODELS FOR SAFTINet

Data modelers work from use cases, which are small vignettes illustrating the tasks an Information System needs to support. A very high-level CER use case is: "Identify a cohort of adult patients with asthma who meet a set of criteria." Table 1 contains a draft cohort definition from the SAFTINet project. Concepts contained in this cohort definition highlight data items that must be present in the SAFTINet data model such as data on patients, diagnoses, medication administrations, filled prescriptions, and emergency department, ambulatory, urgent care, and inpatient visits. The definition also contains explicit or implied constraints that must be represented in the SAFTINet data model: patients must have a date of birth to calculate age; visits must have a date to calculate time intervals; and diagnoses and medication administrations must have a date or be associated with a visit that has a date.

Other use cases revealed additional "must-support" capabilities for the SAFTINet data model, including the ability to:

- extract patient-level data to create analytic datasets;

TABLE 1. A Representative Cohort Definition From the SAFTINet CER Project

1. Adults (ages 18 and over) as of January 1, 2009 receiving care in selected sites who:
Have had at least 2 visits separated by at least 30 d coded as 493.xx in the 18 mo before July 1, 2011, OR
A single diagnosis of 493.xx AND 2 filled prescriptions for an asthma maintenance medication separated by at least 30 d in the past 12 mo
2. Identify the following subcohorts:
$\geq$ 1 asthma exacerbations in the past 12 mo
A visit resulting an oral steroid burst
An emergency department or urgent care visit with code of 493.xx (any position)
$\geq$ 3 outpatient visits within a 14 d period with codes of 493.xx (any position)
$\geq$ 1 hospitalizations related to asthma in the past 12 mo
3. Exclusion criteria: patients with other concomitant chronic lung disease
Cystic fibrosis
COPD, emphysema, chronic bronchitis
α-1-antitrypsin deficiency
Pulmonary fibrosis
Active TB

CER indicates Comparative Effectiveness Research; SAFTINet, Scalable Architecture for Federated Translational Inquiries Network; TB, tuberculosis.

- calculate ages to the year for adults, and to smaller units of measurement for children depending on their age;
- calculate prescribed drug intervals (often called drug exposures);
- link a patient's data across disparate data sources;
- use standardized terminologies to take advantage of conceptual hierarchies and relationships;
- identify a patient as being part of a defined cohort to allow prospective data collection;
- support deidentified data in compliance with HIPAA regulations.

Applications, Platforms, and Data Models

Although a data model is at the core of all large-scale projects, it is not the only consideration in selecting the data management environment for a robust CER infrastructure project. The applications or platforms used by investigators or analysts to access data in a database are also critically important to the overall value of the data infrastructure to end users. The terms "application" and "platform" are often used interchangeably but in our usage, an application completely defines the functions that can be performed by an end-user, whereas a platform is more flexible allowing users with sufficient programming skills to add new functionality to the basic system. Both applications and platforms include software components that interact with the data model. The most visible component is the user interface. The user interface defines which parts of a data model can be "seen" and "manipulated" by the user. A very restrictive user interface can prevent access to a wide range of data model capabilities can make a flexible data model appear to be very limited to the user. Alternatively, a comprehensive set of user

TABLE 2. Definitions for 8 Dimensions of Data Model Quality

Data Model Quality Dimension	Original Definition From Moody and Shanks[28]	Definition Recasted into CER Context
Completeness	Does the data model contain all user requirements?	Can the data model store and retrieve data to meet investigator CER needs?
Integrity	Does the data model conform to the business rules and processes to guarantee data integrity and enforce policies?	Does the data model enforce meaningful data relationships and constraints that uphold the intent of the data's original purpose, that is, clinical care, billing. If required, does the data model recursive relationships to support concept hierarchies?
Flexibility	Does the data model deal with changes in business and/or regulatory change?	Can new data elements and relationships be added if project scope or if regulatory rules (eg, patient identification) change?
Understandability	Are the concepts and structures in the data model easily understood?	Do the concepts, structures, and relationships make sense to investigators, data managers, and statisticians? Are there detailed metadata descriptions to fully describe the correct contexts, intended uses, restrictions, and assumptions for each data element?
Correctness	Does the data model conform to the rules of the data modeling techniques?	Does the model conform to good data modeling practices such as limited data storage redundancy?
Simplicity	Does the data model contain the minimum possible entities and relationships?	Are concepts represented as straightforwardly as possible? Are all data elements necessary? Are the data elements easy to transform from the EMR into the research data model? Are the data elements easy to extract into a research dataset?
Integration	Is the data model consistent with the rest of the organization's data?	Do all of the various data domains, such as demographics, observations, laboratories, and medications "hang together" in a consistent and logical manner? Are tables linked appropriately to meet the research study objectives?
Implementability	Can the data model be implemented within existing time, budget, and technology constraints?	Can the data model be implemented and maintained by current and future partners given anticipated budgets, time, and technical constraints?

CER indicates Comparative Effectiveness Research; EMR, electronic medical record.
From Moody and Shanks.[28]

applications can make a less-than-optimal data model appear more desirable because of the ease-of-use provided by the applications. The selection of a data model, and the resulting database for CER, may or may not depend on the available applications or platforms that support the use of the database. For example, users may need to integrate a data model into an existing platform that provides graphical user interfaces, data visualizations, or statistical software applications.

Build-From-Scratch Versus Adopt-and-Modify

Debates regarding build versus adopt at the initial database design phase are common. Specific decisions are influenced by many factors, chiefly resources and the suitability of available models. The key benefit of build-from-scratch is the opportunity to design the data model with high specificity, essentially creating, within resource limitations, a perfect fit of tool to need. An alternative is to work with an existing data model that could fit user needs with feasible modifications. This approach is particularly beneficial if an active community of users is using the data model and is contributing new software modules or features that enhance the model's value, a benefit from shared investments. A community of users could build-from-scratch a data model that meets multiple needs although models that attempt to meet multiple needs simultaneously add both complexity and compromises. In a similar manner, adopting an existing model also requires adjusting the model to the additional needs of the current project. Both approaches require compromises: the build strategy may be time-limited and resource-limited while the adopt strategy may be constrained by inflexibility from previous design decisions.

In SAFTINet, we were interested in adapting an existing data model rather than defining our own data model. We felt it prudent to build upon prior investments and existing efforts. The ability to begin with a "field-tested" data model and to contribute and expand upon an already established model was determined to be a better investment of SAFTINet resources and that our contributions would provide a return benefit to the original data model's community.

However, there are a number of risks associated with attempting to leverage an existing data model that was designed and optimized for 1 purpose to support a different set of use cases:

1. The candidate data model may store essential data in a manner insufficient for CER use. For example, in Figure 1, both data models require a medication prescription record to be present before drug fulfillment information can be stored. If the CER study depends upon medication fulfillment data that is not tied to prescription data, then both data models in Figure 1 are insufficient to meet this requirement.
2. The existing data model may store data in a manner that is difficult to query for CER. Cohort identification requires concepts such as "average daily exposure period >30 days" or "time between successive admissions." These concepts are not likely to be represented in many data models and would have to be computed either "on the fly" during the query or as a preprocessing step, adding complexity and therefore time and resource to deriving the cohort.
3. The existing data model may have entire data domains absent because they were not relevant to the original project needs. For example, a model created to support drug safety may not contain tables for detailed billing or reimbursement data because these data were not relevant to the original project.

Evaluating an Existing Data Model's Complexity and Usability

The general approach for determining if an existing data model can meet the needs of a new project begins with use cases as described previously. From these use cases, a data analyst examines the proposed data model and determines whether the existing tables and columns in the data model can satisfy the proposed analytic needs. If all of the required data can be stored, the data analyst attempts to estimate the complexity of the queries necessary to extract data from the data model. One key indicator of query complexity is the number of tables that need to be linked to answer a query. For example, in Figure 1, to determine all of the medications that a patient has filled over a period of time would require linking 4 tables in model 1a (Patient → Visit → Prescriptions → Rx Fulfillment) but only 3 tables in model 1b (Patient → Prescriptions → Rx Fulfillment). A more challenging measure of data model fit is assessing the difficulty

TABLE 3. Data Models That Support Observational Comparative Effectiveness Research Considered by SAFTINet

Name	Developing Entity	Initial Purpose
Observational Medical Outcomes Project (OMOP) http://omop.fnih.org/node/35	Foundation of the NIH	Comparative Drug Outcomes Studies
Virtual Data Warehouse (VDW) http://www.kpchr.org/research/public/ourResearchContent.aspx?pageid=90	HMO Research Network	Distributed data warehouse to allow comparative studies across collaborating sites: HMO Research Network, Cancer Research Network & Oregon CTRI
i2b2 http://www.i2b2.org	Partners Healthcare	Scalable informatics framework to investigate disease of genetic origin
OpenMRS https://openmrs.org	Regenstrief Institute	Open source enterprise electronic medical record system platform
OpenEHR http://www.openehr.org/home.html	OpenEHR Foundation	Semantically enabled health computing platform

CTRI indicates Clinical and Translational Research Institute; SAFTINet, Scalable Architecture for Federated Translational Inquiries Network.

TABLE 4. The SAFTINet Criteria for Evaluating Existing Data Models for CER Research

Criteria	Description	CER Investigator Implications	Data Model Technical Implications
Quality Dimension: Completeness			
Data Coverage: The ability to accommodate required CER data elements and domains.	A comparison of the data domains and data elements required in the SAFTINet project that exists in the "out of the box" (supported) version of the data model.	CER domain knowledge and near-final use-cases required	Technical understanding of the data model under review is required to assess data storage options.
Quality Dimension: Flexibility			
Extensibility: The methods used to expand the data model for more data elements, data types and new data domains	Can new elements be incorporated by adding new values to existing data elements (e.g., adding "outpatient" as a new value to VISIT_TYPE_ versus needing to add new tables or columns.	Expanding CER questions and uses	Larger scale domain extensions may require changes to the database technical platform for tuning, indices, and efficiencies
Scalability: Can the model be sized to smaller or larger data sets?	The size of the data sets that have been supported in actual field use.	Domain knowledge regarding anticipated size of our data set and the anticipated rate of growth.	Model size may require underlying infrastructure changes to accommodate.
Adaptability: How broad a variety of data domains may be modeled?	Willingness of existing user community to accept and incorporate data model additions and changes	Active uses group to support sustainability and coverage, synergy of efforts for query development	
Quality Dimension: Understandability			
Understandability: The effort required for technical staff to understand the data model	Similarly, the effort required for data analysts to understand how to construct a CER query against the data model.	Resource staffing consideration, data analyst efficiency	
Quality Dimension: Correctness			
Efficiency:	The growth of the database with absent data (how null values are handled)	Data extraction performance	Infrastructure knowledge is required regarding how nulls may be handled in the database. Null values may change storage requirements if data is exceptionally sparse.
Quality Dimension: Integration			
Use of standardized vocabularies: Does the model support standardized terminologies?		Local terminologies require expertise to map to standards. Standard terminologies permit more ready sharing of data (see grid-"friendly" below.)	
Quality Dimension: Implementability			
Field experience: The number and diversity of uses of the data model. How similar to our use cases are existing uses?	The experiences of users who have attempted to use the database in a manner similar to our intended use.	Willingness of existing user community to engage with SAFTINet during data model investigations.	
Stability: The number of changes to the data model over the past 12-24 months.	Is the data model under review and revised periodically?	A non-stable model requires more action on the part of the business to "keep up" with model changes.	A non-stable model may require underlying infrastructure changes to maintain.
Adoption: The size of the community using and supporting the data model.	This is a reflection of the vibrancy of ongoing data model innovations and data model longevity	The more active the community, the more dynamic the model. Pros include increasing domain and use case accommodation. Cons might include the need for the business to adapt to an oft-changing model.	Technical requirements will be higher to accommodate a dynamic model than a static model. Model changes sometimes necessitate infrastructure changes.
Grid-"friendly": Measure of the model's ability to participate as part of a grid-enabled distributed database	A unique requirement for SAFTINet which seeks to support grid-enabled distributed databases based on a single common data model across disparate clinical and financial data providers.	Success and sustainability of the network, network coverage expansion to support a greater number of CER applications	
Cost: Licensing, staffing, costs of infrastructure	If not public-domain, what are the licensing costs?	Consider staffing and resource costs. Specialized, domain-specific and/or model-specific human resources may be more costly than general report and query development resources.	Consider infrastructure costs: hardware, software, service contracts, Information Technology resources to maintain infrastructure required to run the model.

CER indicates Comparative Effectiveness Research; SAFTINet, Scalable Architecture for Federated Translational Inquiries Network.
Quality Dimensions from Moody and Shanks.[28]

to add a new data element (eg, a new medication) or data class (eg, radiology results).

Data Models Versus Data Quality

A high-quality data model cannot completely ensure high-quality data. A data model that supports the identified use cases can make data collection, recording, and analysis significantly easier and can ensure that entered data meet certain constraints such as data type, allowed values, and mandatory/optional values. But a data model alone cannot ensure that data values make logical or "real-world" sense. For example, a data model can constrain patient sex values to be either "Male" or "Female" and diagnoses to be constrained to valid ICD-9-CM or ICD-10-CM codes but this data model will permit storage of Sex = "Male" and Diagnosis = "V22 Normal Pregnancy." Data quality check, such as male patients cannot be pregnant, require domain knowledge that exists outside of the data model and are data quality processes applied to data that are stored in a data model.

DATA MODELS CONSIDERED BY SAFTINet

Table 3 lists the data models considered by SAFTINet during our requirements analysis phase. An absolute requirement for all candidate models was that the model be freely available in the public domain without licensing restrictions. An equally important consideration was the existence of an active user community that the SAFTINet team could leverage for advice, guidance, and collaboration.

The 8 quality dimensions in the Moody and Shanks framework in Table 2 are described at a high level of abstraction. The SAFTINet team expanded the general framework descriptions into more detailed project-specific criteria for 6 of the Moody and Shanks dimensions (Table 4). The dimensions and specific criteria were developed iteratively as the SAFTINet team explored more detailed use cases and requirements for the SAFTINet project.

Although all of the data models listed in Table 3 meet many of the SAFTINet criteria in Table 4, the team selected the Observation Medical Outcomes Partnership (OMOP) data model. Specific technical design features of the OMOP data model allow for a broad range of clinical observations to be added without any structural changes to the model (no new tables or columns), handle missing data without creating empty cells, and have extensive field testing with very large administrative and clinical datasets that support complex analytic methods. OMOP exploits a rich set of terminology tables to simplify complex queries involving conceptual groups (eg, "antibiotics"). In addition, the OMOP public web site contained extensive metadata documentation and examples to clarify the intended use of each table and field. And the timing was right—the OMOP team was actively seeking new collaborators to extend the current OMOP data model to support new areas of outcomes research beyond its initial focus on drug surveillance. The SAFTINet CER network is currently being constructed with OMOP Common Data Model Version 3.0 (see http://omop.fnih.org/CDMV3 for the full specifications).

DISCUSSION

Ensuring that a database can store and retrieve data required to create analytic datasets begins with an understanding of the anticipated data uses. Development of detailed use cases that describe the research hypotheses, cohort selection criteria, and analytic plans that the database must support requires the active engagement of clinical investigators. Because clinical research and related activities are both hypothesis driven and exploratory, it is often difficult for investigators to delineate all intended uses of a database. However, the urge to "keep all options open" will result in a data model that is overly complex to use and maintain, manifesting in a proliferation of tables and linkages that must be navigated when storing and retrieving data from the database.

Clinical investigators typically focus on the structure and content of analytic datasets. Less attention is given to the structure and content of supporting databases where data from disparate data sources are combined, integrated, and harmonized. Small studies with limited data drawn from a single data source do not have these issues. But as a study expands in size, scope, locations, data sources, and investigator community, the lack of attention to the data model can bring a previously successful pilot study to a grinding halt. With CER's focus on a wide range of data from different sources across broad patient populations, the diversity of data types that need to be accommodated in research databases has grown significantly. For example, although SAFTINet does not have a requirement to represent primary "-omics" or sequence data, the OMOP model is robust enough to accommodate such data in future.

Investigators may attempt to create a data model from scratch. For a very large project that will be a reusable resource for multiple investigations, this approach should be considered very carefully. Existing data models that have been field tested and have an active user community represent hundreds of hours of analysis and use that have revealed strengths and weaknesses.

Even with access to experienced data modeling skills, the SAFTINet team made an early decision to not develop a new data model, recognizing that adopting an existing data model would require adaptation. We selected a model that could accommodate anticipated changes without major rework. The presence of a large active user community and a development staff willing to help extend the model in a manner congruent with the original model were also strong determinants. Other models listed in Table 3 have similar characteristics and would be equally valid selections depending on requirement prioritization (Table 4). By selecting an existing data model, SAFTINet will contribute back to the user community with additional use cases and added functionality. In this respect, SAFTINet will be gaining from and contributing to the work of other CER investigators and data modelers.

REFERENCES

1. Blumenthal D. Launching HITECH. *N Engl J Med*. 2010;362:382–385.
2. Waldren S, Kibbe DC, Mitchell J. "Will the feds really buy me an EHR?" and other commonly asked questions about the HITECH Act. *Fam Pract Manag*. 2009;16:19–23.
3. McLeod A. Health IT for economic and clinical health: HITECH Medicare incentive payment estimator. *Healthc Financ Manage*. 2009;63:110–111.

4. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med*. 2010;2:57cm29–31.
5. Grossmann C, Institute of Medicine (US). *Roundtable on Value & Science-Driven Health Care. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting A Public Good: Workshop Summary*. Washington, DC: National Academies Press; 2010.
6. Agency for Healthcare Quality and Research. Expansion of the DEcIDE distributed research network (DRN) infrastructure to support studies of comparative effectiveness—Phase II. *Effective Health Care Program* 2010. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=485. Accessed August 30, 2011.
7. Bekke JHt. *Semantic Data Modeling*. New York: Prentice Hall; 1992.
8. Simsion GC. *Data Modeling: Theory and Practice*. Bradley Beach, NJ: Technics Publications; 2007.
9. Hoberman S. *Data Modeling Made Simple: A Practical Guide for Business and IT Professionals*. 2nd ed. Bradley Beach, NJ: Technics Publications; 2009.
10. Borkin SA. *Data Models: A Semantic Approach for Database Systems*. Cambridge, Mass: MIT Press; 1980.
11. Chmura A, Heumann JM. *Logical Data Modeling: What it is and How to do it*. New York, NY: Springer; 2005.
12. Tsichritzis DC, Lochovsky FH. *Data Models*. Englewood Cliffs, NJ: Prentice-Hall; 1982.
13. Silverston L. *The Data Model Resource Book*. Revised ed. New York: John Wiley; 2001.
14. Riben M, Wade G, Edgerton M, et al. Aligning tissue banking data models for caBIG interoperability. *AMIA Annu Symp Proc*. 2008;1109.
15. Brown JS, Lane K, Moore K, et al. *Defining and Evaluating Possible Database Models to Implement the FDA Sentinel Initiative*. U.S. Food and Drug Administration; 2009. Available at: http://www.regulations.gov/contentStreamer?objectId=0900006480e6cd93&disposition=attachment&contentType=pdf. Accessed May 1, 2012.
16. Marrs KA, Kahn MG. Extending a clinical repository to include multiple sites. *Proc Annu Symp Comput Appl Med Care*. 1995;387–391.
17. Marrs KA, Steib SA, Abrams CA, et al. Unifying heterogeneous distributed clinical data in a relational database. *Proc Annu Symp Comput Appl Med Care*. 1993;644–648.
18. Krohn R. Advice on HIE for the ARRA-minded. A big boost for digital transformation. *J Healthc Inf Manag*. 2009;23:7–8.
19. Halamka JD. Making the most of federal health information technology regulations. *Health Aff. (Millwood)*. 2010;29:596–600.
20. von Halle B. Data: asset or liability? *Database Program Des*. 1991;4:7–9.
21. Batini C, Ceri S, Navathe S. Conceptual database design: an entity-relationship approach. *Redwood City, CA*. Benjamin/Cummings Pub. Co; 1992.
22. Levitin A, Redman T. Quality dimensions of a conceptual view. *Inform Process Manag*. 1995;31:81–88.
23. Krogstie J, Lindland O, Sindre G, et al. Towards a deeper understanding of quality in requirements engineering. *Adv Inf Syst Eng*. 1995;932:82–95. Springer Berlin/Heidelberg.
24. Lindland OI, Sindre G, Solvberg A. Understanding quality in conceptual modeling. *IEEE Software*. 1994;11:42–49.
25. Kesh S. Evaluating the quality of entity relationship models. *Infor Software Technol*. 1995;37:681–689.
26. Simsion GC, Witt GC. *Data Modeling Essentials*. 3rd ed. Amsterdam; Boston: Morgan Kaufmann Publishers; 2005.
27. Moody DL, Shanks GG. What makes a good data model? Evaluating the quality of entity relationship models. *Proceedings of the13th International Conference on the Entity-Relationship Approach*; 1994.
28. Moody DL, Shanks GG. Improving the quality of data models: Empirical validation of a quality management framework. *Inf Syst*. 2003;28:619–650.
29. Moody DL. Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data Knowl Eng*. 2005;55:243–276.

