Desiderata for Healthcare Integrated Data Repositories Based on Architectural Comparison of Three Public Repositories

Vojtech Huser, MD, PhD, James J. Cimino, MD
Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD

Abstract

Integrated data repositories (IDRs) are indispensable tools for numerous biomedical research studies. We compare three large IDRs (Informatics for Integrating Biology and the Bedside (i2b2), HMO Research Network's Virtual Data Warehouse (VDW) and Observational Medical Outcomes Partnership (OMOP) repository) in order to identify common architectural features that enable efficient storage and organization of large amounts of clinical data. We define three high-level classes of underlying data storage models and we analyze each repository using this classification. We look at how a set of sample facts is represented in each repository and conclude with a list of desiderata for IDRs that deal with the information storage model, terminology model, data integration and value-sets management.

Introduction

Current clinical and translational research increasingly relies on the existence of robust integrated data repositories (IDRs) with administrative, clinical, and "-omics" data.[1] Following clear warehouse design principles can lower long-term maintenance costs for organizations that are currently building or significantly restructuring their data warehouses. Maintenance of those warehouses is very costly, and architectural changes are complicated by existing dependencies. Getting the right architecture early during the warehouse creation is crucial. We set out to compare three IDR architectures in order to identify common architectural features and advantages and formulate a list of desiderata. The objective to provide an integrated data repository to researchers, clinicians, and administrators can be met in number of ways. However, our prior experience, and that of others, shows that adhering to certain principles leads to a more robust design that is better able to meet current known and future unforeseen requirements. We claim that formulating a set of requirements for a data warehouse may prove similarly beneficial as was formulation of desiderata for controlled terminologies.[2]

To decide which parameters to compare and on which to focus, we considered existing prior literature about IDRs: Huff formalized an event-based model for organizing individual facts stored in an IDR;[3] Murphy described several optimizations for relational databases;[4] Nadkarni offered an extensive account on database design[5] and Gilchrist looked at query speed optimizations.[6] Also relevant are properties of informatics platforms for conducting comparative effectiveness research (CER) as analyzed by Sittig.[7]

Background

IDR types: We considered two high-level types of IDRs when deciding which repositories to analyze and compare: (1) a *single institution schema* (or a single vendor) that captures a large number of possible data domains, and (2) an *integrative IDR schema* that strives to capture a limited set of common data domains from multiple institutions. Examples of a single institution warehouse are those of Intermountain Healthcare, Partners HealthCare or Regenstrief Institute. At such institutions, often with homegrown EHR systems, the importance of having a data warehouse is well understood and there are decades of experience with data warehouse evolution. However, the structure of such warehouses is often not published in detail, and, in some cases, completely inaccessible due to copyright protection. On the contrary, integrative IDR schemas often make their structure publicly available in order to promote adoption. Integrative IDRs also tend to be less complex, and have fewer data tables in order to focus on common data domains of multiple institutions (as opposed to storing all possible data at a given individual site). For reasons of complexity and public availability, we chose to focus on analyzing the architectures of the integrative IDRs.

Schema models: For the overall characterization of the warehouses, we defined three high-level data organization models: (1) An *entity-attribute-value (EAV) model* that stores several attributes in a more generic table (e.g., both laboratory results and procedure events would be fact instances stored in a single data structure). This principle can also be applied to additional details about a fact (sometimes called attributes, modifiers or parameters). Furthermore,

the attribute principle can be applied at single as well as multiple layers. E.g., each instance of an EAV-based event table (e.g., biopsy event) may have many event attributes (who ordered the biopsy) stored in an associated attribute-EAV-based table. (2) A *hybrid model* stores some elements in an EAV mode but certain common event attributes have a designated column (e.g., fact_source_system, observation_type or observation_value_text). Providing data for such hard-coded columns may not be required and they remain empty for some facts.[8] Often event_time is one such attribute and an EAV model is sometimes extended to an entity-attribute-value-time (EAVT) model. (3) A *traditional model* (column-based) stores each subset of data (e.g., encounters, procedures or oncologic attributes) in specialized tables with columns representing necessary fact attributes (e.g., tumor table with tumor stage and tumor type columns).

IDR requirements: For final formulation of desiderata, we assumed the following basic requirements for an IDR:

- *Re-use:* routine care data are re-used for research purposes or clinical purposes (e.g., inform care of patients based on past experience with similar patients)
- *Integration:* data are integrated to facilitate long-term lifetime analysis, such that data from disparate sources are linked to the corresponding patient (billing data, clinical data) and linked to the corresponding event (order entry, order fulfillment). Moreover, semantically identical or semantically related data are also linked.
- *Organization:* the IDR can accommodate a wide range of source systems and is easy to use and extend. It strikes a balance between graceful evolution and stability of the data structures. For example, major restructuring does not occur often and most new data sources can be integrated without major schema change
- *Maintenance:* the IDR is optimized for easy maintenance, especially with respect to adapting to changes in source systems, and is robust to turnover of maintenance staff and data analysts

Methods

Sample selection

We initially considered a large set of IDR architectures published in the informatics literature that included architectures of Informatics for Integrating Biology and the Bedside (i2b2),[9,10] HMO Research Network's Virtual Data Warehouse (VDW),[11,12] the Observational Medical Outcomes Partnership (OMOP),[13] DARTNet,[14] HealthFlow,[15] and repositories at Intermountain Healthcare,[3] NIH,[16] Mayo Clinic,[17] Stanford University,[18] Columbia University,[19] Duke University[20] and others listed on an IDR research community wiki.[21] For final analysis, we chose a purposive sample of IDRs for which detailed schema documentation is available and IDRs that are of integrative type rather than single institution IDRs. The three finally selected architectures were i2b2, VDW, and OMOP. Diagrams of the three analyzed IDR schemas (with links to their documentation) are available at the project website at http://code.google.com/p/desiderata.

Comparison methodology

Prior studies in knowledge representation of coded healthcare data clearly describe a close relationship between an information model for storing facts and the employed terminology model.[3] Because of this close relationship, we analyze the architectures in two aspects: (1) architecture for storing facts, as well as (2) structures for representing the terminology layer of the warehouse. By *terminology layer* we refer to parts of the IDR architecture that deal with representation of coded medical concepts (e.g., diseases, medications, laboratory findings and diagnostic procedures, and visit types) separately from structures representing individual patients' clinical facts (e.g., Patient John Doe's diagnosis of Parkinson's disease on Sep 23, 2012). The terminology layer may simply be a collection of external terminologies, such as ICD-9-CM, LOINC or RxNorm, but in many cases it includes a comprehensive IDR internal terminology, that we refer to as a *native terminology*, with locally defined terms to support semantic data integration. An example of native terminology is the Medical Entities Dictionary[22] at Columbia University, or the Healthcare Data Dictionary at Intermountain Healthcare.[23]

To add a more practical level of insight, and to be able to make analogous comparison across the three analyzed IDRs, we look at how each repository would store a set of five sample events in addition to abstract architectural analysis. The first two sample events targeted representation of currently common data, while the remaining three targeted storage of emerging and recently suggested data domains (later referred to as novel data types). The sample events were: (1) representing a laboratory result; (2) storing occurrence of a particular healthcare procedure; (3) storage of data from electronic case report forms (eCRF) collected during patient's participation in a clinical trial;

(4) storage of results of pharmacogenetic tests, such as Affymetrix DMET™ genotyping array; and (5) storage of structured family history data (e.g., maternal grandfather of patient X died of melanoma at age 36).

Results

Schema comparison

Table 1 shows an overview of the high-level repository parameters we analyzed in this study. The three compared IDRs differed in their approach to various well-established or less-common data domains (e.g., diagnoses, lab results, or medications). For example, the VDW repository defines separate tables for lab results, vital signs and tumor facts, while the i2b2 and the OMOP repositories use a generic table approach that can accommodate multiple data domains. Considering the degree of adoption of the EAV paradigm and a generic fact table, i2b2 is the biggest adopter since it uses the observation fact for all data domains, while OMOP still separates data domains of diagnoses, procedures or medications from their generic observation table (see 'Generic Fact Structure' and 'Designated Data Structures' rows in Table 1).

Table 1. High-level architectural comparison of the three analyzed IDRs

PROPERTY	i2b2	OMOP	VDW
Generic fact data structure	OBSERVATION_FACT	OBSERVATION	n/a
Designated data structures	PATIENT_DIMENSION, VISIT_DIMENSION, PROVIDER_DIMENSION	PERSON, VISIT_OCCURENCE, DEATH, COHORT, PROVIDER, CARE_SITE DRUG_ERA, DRUG_EXPOSURE, CONDITION_ERA, CONDITION_OCCURENCE, PROCEDURE_OCCURENCE	DEMOGRAPHICS, CENSUS, ENCOUNTERS, ENROLLMENT, DEATH, PROVIDER, VITAL SIGNS, LAB_RESULTS, DIAGNOSES, PROCEDURES, PHARMACY, TUMOR
Terminology layer	CONCEPT_DIMENSION	CONCEPT, CONCEPT_RELATIONSHIP, CONCEPT_ANCESTOR, SOURCE_TO CONCEPT_MAP	No generic terminology table; EVER_NDC table (for drug codes only)
Fact nesting	Generic *modifier_cd* column (coded in native terminology) in the OBSERVATION_FACT table	Generic *obs_value_as_concept_id* column (coded in native terminology) in the OBSERVATION table. Domain-specific columns in designated tables. Additional fact grouping (temporal, functional) via PAYER_PLAN_PERIOD table and several _ERA tables.	No generic fact nesting structure. Numerous domain-specific columns in designated tables (e.g., encounter type in PROCEDURES). Additional fact grouping (temporal) via ENROLLMENT table.
Designated columns in fact table	valtype_cd, units_cd, encounter_num, provider_id, location, confidence_num, valueflag_cd, observation_blob	observation_type_concept_id, obs_range_low, obs_range_high, associated_provider_id, source_obs_code, unit_concept_id	n/a

The warehouses also differ in the degree of complexity of their terminology layer. The VDW repository has a formal native terminology for medications but not for other domains. Instead, it uses individual tables' metadata specifications to define coded values and corresponding meaning for data in many VDW tables and columns (see 'Terminology Layer' row in Table 1). The i2b2 and OMOP repositories do include a formal terminology layer but differ in how concepts can be hierarchically grouped together. Tables 2 and 3 show OMOP and i2b2 representation of three example procedure leaf concepts ("chest wall incision", "thoracotomy" and "pleuroperitoneal shunt creation") together with two parent concepts ("operations on chest wall, pleura, mediastinum and diaphragm" and "operations on respiratory system"). Both IDRs technically support multiple hierarchies (a terminology concept can have multiple parent concepts). The OMOP repository uses a more elaborate structure with a CONCEPT_RELATIONSHIP table (see Table 2) that supports different relationships (e.g., is_a/subsumes [inverse is_a], has_ingredient/is_ingredient_of, has_severity/severity_of, concept_replaced_by/concept_replaces) and a CONCEPT_ANCESTOR table that can be used to obtain all child concepts (direct and inferred via the is_a relationship). The i2b2 terminology uses a single CONCEPT_DIMENSION table (Table 3) and relies on a "concept path" column to organize concepts into hierarchies. i2b2 also implements only a single relationship type between any two concepts.

Table 2: OMOP's CONCEPT_RELATIONSHIP table example showing sample concepts.

CONCEPT_1	RELATIONSHIP*	CONCEPT_2
Incision of chest wall	is_a	Incision of chest wall and pleura
Exploratory thoracotomy	is_a	Incision of chest wall and pleura
Creation of pleuroperitoneal shunt	is_a	Incision of chest wall and pleura
Incision of chest wall and pleura	is_a	Operations on chest wall, pleura, mediastinum, and diaphragm
Operations on chest wall, pleura, mediastinum, and diaphragm	is_a	Operations on respiratory system

*The reverse relationships ("subsumes") are not shown. Relationship is shown directly as a description rather than as relationship ID.

Table 3: i2b2's CONCEPT_DIMENSION table example showing sample concepts.

CNCPT_CD	CONCEPT_PATH	NAME_CHAR
ICD9:34.01	\i2b2\Proc\Operations on respiratory system\Operations on chest wall, p~\ Incision of chest wall an~\Incision of chest wall\	Incision of chest wall
ICD9:34.02	\i2b2\Proc\Operations on respiratory system\Operations on chest wall, p~\ Incision of chest wall an~\Exploratory thoracotomy\	Exploratory thoracotomy
ICD9:34.05	\i2b2\Proc\Operations on respiratory system\Operations on chest wall, p~\ Incision of chest wall an~\Creation of pleuroperito~\	Creation of pleuroperitoneal shunt
ICD9:34.0	\i2b2\Proc\Operations on respiratory system\Operations on chest wall, p~\ Incision of chest wall an~\	Incision of chest wall and pleura
ICD9:34	\i2b2\Proc\Operations on respiratory system\Operations on chest wall, p~\	Operations on chest wall, pleura, mediastinum, and diaphragm

To better characterize and describe the repository's information model, we define a term *fact nesting* to refer to the ability of the IDR schema to represent one or more *nested facts* (or attributes) related to a single *master fact* (or master event). Storing nested facts may employ the use of an event ID mechanism to properly differentiate which nested facts extend which master events.[16] In many IDRs, single-level fact nesting is achieved by additional fact table columns, such as modifier code, without using an event ID mechanism. Examples of fact nesting are: (1) a microbiology result with antimicrobial susceptibility testing sub-results; or (2) an order set, such as an admission order set, with several component orders. The VDW schema provides one or several pre-defined nested fact columns depending on the data domain, whereas the i2b2 and OMOP schemas both include a generic second attribute column (modifier_cd in i2b2) or second value column (obs_value_as_concept_id) that could be directly utilized or overloaded for fact nesting. Despite the existence of such generic fact nesting structures, common event attributes often have designated columns defined in the repository schema, such as location, observation type, fact source, associated provider (see the last row in Table 1).

Analysis of individual repositories

i2b2: The key data structure is the observation_fact table. The attribute column is called concept_cd and the table also includes columns for text value, numerical value and flag value. Modifier_cd is used to store additional codes related to the main attribute and implements a single level fact nesting structure. The observation_fact table is technically a hybrid EAV table with several hardcoded columns (e,g, location, confidence or units). The schema contains a few other tables (PATIENT_DIMENSION, VISIT_DIMENSION, PROVIDER_DIMENSION) using a traditional modeling approach. Additional architectural constraints are defined for institutions federating multiple i2b2 repositories as a SHRINE network.[24]
Terminology: The schema includes an explicit model of a terminology layer in a concept_dimension table. A single concept_path column is used to model inter-term relationships.
Example data: Using our data storage examples, laboratory values and procedure events would both be stored in the observation_fact table. The observation_fact table is capable of storing new data domains given a proper prior terminology representation and most facts from the novel data scenarios could be stored in an i2b2 repository.

OMOP: The schema contains designated tables for the domains of diagnoses, procedures and medications. It includes a somewhat generic table called OBSERVATION with three versatile columns for numeric data (obs_value_as_number), textual data (obs_value_as_string) and coded value data (obs_value_as_concept); however,

the table appears to be optimized for storing laboratory values (by presence of columns for high and low observation range).
Terminology: The schema includes a complex terminology layer with support for multiple hierarchies, multiple relationships (in addition to a default "is_a" relationship), and synonyms. The warehouse documentation includes code snippets for common terminology questions (e.g., "what are all parent concepts for a given concept", or "list all drugs that have same indications as drug X").
Example data: Laboratory values would be stored in the OBSERVATION table. Procedures would be stored in the PROCEDURE_OCCURANCE table. Novel data types could be stored in the OBSERVATION table by using the obs_value_as_concept column and the native terminology.

VDW: The schema defines 11 fact tables, each covering a well-defined data domain (e.g., VITAL_SIGNS, PROCEDURES, or TUMOR). They follow either the hybrid EAV model (e.g., PROCEDURES, DIAGNOSES) or the traditional model (e.g., DEMOGRAPHICS, CENSUS, VITAL_SIGNS). VDW specifications include extensive documentation at the table and column level. The HMO Research Network maintains a library of data quality assessments code snippets that can compare data patterns across sites.[25]
Terminology: The schema does not include an explicit terminology layer, except for the EVER_NDC table storing medications codes. A separate Provider table stores specialties of clinicians.
Example data: Laboratory values would be stored in the LAB_RESULTS table. Procedures would be stored in the PROCEDURES table. Considering the existing underlying VDW modeling approach, novel data types would most likely be stored in a new table.

IDR comparison results

The comparison of the available documentation of all three IDRs showed several recurring themes. The following features were found in all three compared repositories: (1) use of EAV structure for at least one data domain; (2) use of a single patient identifier with an identical column name across all tables; (3) use of internal terminology layer for at least one data domain; (4) use of an encounter ID to group events relevant to a single healthcare encounter; (5) presence of structures representing facts not related to patients but organizational or regional context knowledge, such as provider data; (6) representation of demographic data in a traditional table, despite the ability to treat those as patient entity attributes within an EAV-based data model. Additional features common to at least two repositories were: (7) a separate death data table (similarly to demographics data domain) (VDW, OMOP); (8) an elaborate native terminology layer with ability to maintain domain-specific value set knowledge (e.g., encounter types, medication administration route) (i2b2, OMOP); (9) a separate table capturing history of patient's specific insurance plan (VDW, OMOP)

IDR Desiderata

Based on the above analysis of three IDRs, as well as our close experience with additional warehouses (NIH's BTRIS, HealthFlow,[15] and the Columbia University IDR) and review of published IDR literature, we formulate a set of desirable characteristics, or "desiderata" for a generic IDR. We analyzed features that lead to positive long-term benefits, and looked how often they are implemented by various IDRs with the goal of formulating our list. We used a combination of (1) a bottom-up approach, where we generalized from features present in various IDRs, and (2) a top-down approach, where we considered general IDR requirements presented in the background section. Some of the desiderata are specifically relevant to a health care data warehouse, while others are more general and applicable to any data warehouse; we include the latter here because they are of particular relevance to healthcare IDRs.

1. **Single patient identifier (ID) and patient ID management:** The IDR should use a single patient identifier in all domains within the IDR. If multiple systems are integrated where different patient IDs are used, an enterprise master patient index[26] should be used to merge corresponding records. To facilitate research analysis, the warehouse should also have a clear model for shadow ID management. A *shadow ID* is defined as a project-specific replacement ID for either patient ID or other identifier within exported or displayed data. The most common shadow ID is a substitute for the patient ID; however, obfuscating provider ID or facility ID is also common. The process of generating shadow IDs must sometimes include ability to re-identify the same patient, if additional data are later requested. HIPAA law mandates keeping record of how each patient's record was used by any relevant research project. At other times, the complete opposite (inability to re-identify) is requested and different shadow ID management techniques (such as discarding the encryption key) are used for that. Some warehouse architectures include a build-in static shadow ID that may be used for one-time views of

data. An IDR should have a process in place for managing and documenting generation of shadow IDs used for all individual data extract (at a single project level) or extracts for a given research group (at a principal investigator team level), or other formal approach.

2. **Information storage model:** An IDR should formally define its information model[3] for storing facts. This information model should be sufficiently generic, extensible and relatively stable in time, so that new data sources can be integrated without major changes to the IDR schema. A sufficiently characterized information model implies existence of documentation that clearly states the purpose of all crucial fact tables and describes a general extract, transform and load (ETL) strategy for integration of new data sources into the existing IDR schema.

3. **Support for fact nesting:** Although IDR data integration by definition involves a significant degree of transformation of original data, an IDR should offer storage structures that allow preservation of how groups of related facts relate to each other. An information model should define what level of fact nesting is possible and define explicitly how nested facts can be linked to master events.

4. **Semantic integration:** Whereas the single patient ID achieves technical data integration, to facilitate research analysis, data should be semantically integrated as well. The IDR should use coded concepts to organize and integrate facts in data domains from related sources. The need for semantic integration is most apparent when an IDR receives similar data from two different sources, such as inpatient and outpatient pharmacy dispensing systems, or legacy and current laboratory information systems. However, it can also arise when integrating multiple related data subdomains from a single source system.[27] Semantic data integration can be achieved by maintaining direct mapping to custom concepts or it can be shifted to the terminology layer.

5. **Terminology model:** Historic developments in IDRs show that large warehouses often include a native terminology layer[22] and that strict reliance on only external terminologies is not sufficiently flexible. As outlined above in the Methods section, the terminology layer often consists of a native IDR terminology that represents a collection of terms that are defined locally by the repository. The native terminology may address: (1) terms that are not defined by any of the data contributing system (supporting the general infrastructure of the IDR); (2) terms that are meant to integrate disparate limited-scope terminologies within the individual data sources (e.g., two appointment scheduling systems) or (3) reconcile semantically two or more related external terminologies (as implied by the semantic integration desideratum). Often challenging is the relationship of the native terminology (or the IDR terminology layer in general) to large, mainstream terminologies, such as SNOMED, ICD, RxNorm, or LOINC (e.g., a strategy for full integration into the native terminology, a strategy for explicit exclusion, or some other approach).[28] The native terminology layer may also play a key role in loading new data into the IDR. A *terminology driven ETL process*, for example, can automate some steps for adding new terms and make the changes consistent and transparent. Such a process automatically detects the presence of new terms in incoming data that lack corresponding formal concepts in the native terminology. We recognize, however, that native IDR terminology can be costly to build initially and maintain later, and that in some data integration efforts, strict reliance on external terminologies is sufficient.

6. **IDR context representation:** Proper knowledge of context is important for accurate data analysis. An IDR may need to be able to represent contextual information on multiple levels, in addition to storing individual clinical facts, such as contextual data about the healthcare organization itself, data about individual medical facilities, and current and past informatics systems or providers. For example, false conclusions about care may be drawn from IDR data that lacks any dialysis events in chronic kidney disease patients simply because the outpatient integrated delivery network contributing the data does not own any dialysis centers. Individual facility data, such as absence of magnetic resonance imaging (MRI) at a given rural location, can similarly bias quality of care or other analyses. The need for context data on providers is exemplified by VDW's PROVIDER table with data on provider's specialties.

7. **Documentation and metadata:** Good documentation[17] of table and column structures as well as ETL processes greatly facilitates correctly formed queries or query speed optimizations (indication of presence of indexes). Metadata often include information that goes beyond the documentation implied by the prior desiderata of the information storage model and terminology model. A few examples include: documentation on when a table was created, ETL processes impacting the table, and an up-to-date link to the human custodian that is most knowledgeable about the structure of and data in each table. Wiki-like metadata documentation can even support creation of community-created comments and transfer of knowledge from IDR staff to or between analysts. An accurate metadata knowledge base is important during IDR staff changes or for training of new

analysts. A code snippet library of common analytical tasks (e.g., obtaining all children terminology concepts of a given term) may be an optional part of a metadata platform.

8. **Capture IDR historical evolution:** IDRs integrate several disparate systems for various domains. Documenting individual systems and relevant milestone dates is important for later data analyses and avoiding artificial, false data patterns. Examples of historical evolutions that should be well documented are: (1) system X was implemented on date D and legacy systems used prior system X have not been integrated in the IDR;(2) system Y to manage radiology reports was replaced by system Z with the following roll-out scheme across different regions within an integrated healthcare delivery network. The purpose of the warehouse is often to provide a long-term data view cutting across current and legacy systems. Proper documentation of historical evolution is important for unbiased data analysis and is superior to detecting systems transition via data reverse-engineering. Important IDR metadata is often lost when a key IDR staff member leaves the organization; proper historical evolution documentation may limit resulting inefficiencies.

9. **Protected Health Information management:** Due to a common request to produce HIPAA "limited data sets", a general IDR strategy that identifies all data elements that may contain PHI facilitate data export. In addition to PHI pertinent to the patient, IDR facts may include attributes about third parties (e.g., provider ID, procedure technician ID) that may also have to be omitted in certain data views and exports. Hence a hierarchical list distinguishing several levels is often justified. In addition, a different approach is needed to handle PHI in textual clinical reports.

Discussion

There are far too many concurrent IDR efforts to review in this paper. The choice of the three selected IDRs was mainly dictated by our ability to have sufficiently detailed public information about their architecture. Our list of desiderata is informed by this limited analysis and extended by the authors' knowledge of other IDRs. This list of desiderata is not intended to be complete, but rather should serve to facilitate discussion about additional desirable characteristics. IDRs may have valid reasons not to adopt a particular requirement; however, we believe that a general philosophy of adherence can prove highly beneficial in long-term IDR maintenance, despite significantly larger initial investment of various resources.

Our study has several limitations. First, we used a limited and purposeful set of repositories to compare; however, public availability and/or copyright protection were the main limiting factors and those were outside of our control. Second, there was a varying degree of available documentation about the eventually selected IDRs and some detailed aspects could not be comprehensively compared. VDW and i2b2 provide discussion platforms (e.g., listservs and wikis) that can be helpful in clarifying some detailed modeling aspects. The VDW's discussion platform, however, is not public and can only be accessed by HMO Research Network members. Finally, our analysis was limited to the relational database paradigm; however, all analyzed repositories and vast majority of healthcare warehouses use a relational database. Various types of emerging schema-free noSQL databases may offer additional findings.

In our final list of desiderata, it was often difficult to draw boundaries between individual requirements and our division may be subjective to some degree. For example, "Support for fact nesting" can be easily viewed as part of "Information model"; similarly, "Capture IDR historical evolution" might be subsumed by "Documentation and metadata". We also omitted some issues, such as (1) regular and computational data quality assurance; (2) dealing with limited data clean-ups versus keeping all data homogenously inconsistent, (3) defining clear boundaries when native terminology concepts should be formally created and (4) providing multiple data access modalities (query tools, human mediated queries, command-line based data access via an API).[29]

Conclusion

In an architectural comparison of three IDRs, we described several features that are common and beneficial in storing and organizing clinical data. Based on this review, our list of IDR desiderata offers advice to institutions newly creating or restructuring their IDRs. Whereas the initial design of many clinical data repositories was driven by provision of decision support or evaluation of quality of care, their research use is rapidly increasing with significant impact on their design. As with similar efforts in informatics, adherence to general principles will provide some immediate benefits, with the potential for future, unanticipated benefits as well.

Acknowledgments and Disclaimers: This work has been supported by intramural research funds from the NIH Clinical Center and the National Library of Medicine. Mention of particular commercial products does not imply endorsement of those products.

References

1. Mackenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc. 2012 Jun 1;19(e1):e119-e24.
2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998 Nov;37(4-5):394-403.
3. Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An event model of medical information representation. J Am Med Inform Assoc. 1995 Mar-Apr;2(2):116-34.
4. Murphy SN, Morgan MM, Barnett GO, Chueh HC. Optimizing healthcare research data warehouse design through past COSTAR query analysis. Proc AMIA Symp. 1999:892-6.
5. Nadkarni PM. Metadata-driven software systems in biomedicine. New York: Springer; 2011.
6. Gilchrist J, Frize M, Ennett CM, Bariciak E. Performance Evaluation of Various Storage Formats for Clinical Data Repositories. Instrumentation and Measurement, IEEE Transactions on. 2011;60(10):3244-52.
7. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care. 2012 Jul;50 Suppl:S49-59.
8. Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. J Am Med Inform Assoc. 2003 Sep-Oct;10(5):444-53.
9. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc. 2007:548-52.
10. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):181-5.
11. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. J Natl Cancer Inst Monogr. 2005(35):12-25.
12. Chapter 4: Virtual Data Warehouse, Collaboration toolkit. Available from: http://www.hmoresearchnetwork.org/resources/toolkit/HMORN_CollaborationToolkit.pdf#4.
13. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600-6.
14. Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009 Sep 1;151(5):338-40.
15. Huser V, Rasmussen LV, Oberg R, Starren JB. Implementation of workflow engine technology to deliver basic clinical decision support functionality. BMC Med Res Methodol. 2011;11:43.
16. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform. 2010;160(Pt 2):1299-303.
17. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc. 2010 Mar-Apr;17(2):131-5.
18. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc. 2009;2009:391-5.
19. Wilcox AB, Vawdrey DK, Chen YH, Forman B, Hripcsak G. The evolving use of a clinical data repository: facilitating data access within an electronic medical record. AMIA Annu Symp Proc. 2009;2009:701-5.
20. Horvath MM, Winfield S, Evans S, Slopek S, Shang H, Ferranti J. The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement. Journal of biomedical informatics. 2011 Apr;44(2):266-76.
21. Clinical Integrated Data Repositories. Available from: http://clinfowiki.org/wiki/index.php/CIDR.
22. Baorto D, Li L, Cimino JJ. Practical experience with the maintenance and auditing of a large medical ontology. Journal of biomedical informatics. 2009 Jun;42(3):494-503.

23. Clayton PD, Narus SP, Huff SM, Pryor TA, Haug PJ, Larkin T, et al. Building a comprehensive clinical information system from components. The approach at Intermountain Health Care. Methods Inf Med. 2003;42(1):1-7.
24. Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc. 2012 Jun 25.
25. Bachman D, Field T, Bredfeldt C, Hornbrook M, Bauck A, Tavel H, et al. PS2-51: Utilization Quality Assurance: Are We Better Yet? Clin Med Res. 2012 Aug;10(3):195-6.
26. Fernandes L, Brandt M, Fletcher D, Grant K, Hatton L, Postal S, et al. Building an enterprise master person index. J AHIMA. 2004 Jan;75(1):56A-D.
27. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. AMIA Annu Symp Proc. 2011;2011:1454-63.
28. Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, et al. Implementation of a deidentified federated data network for population-based cohort discovery. J Am Med Inform Assoc. 2012 Jun 1;19(e1):e60-e7.
29. Integrating R efficiently to allow secure, interactive analysis within a clinical data warehouse, Use R! Conference Proceedings, 2012. Available from: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/UseR-2012/141-Connolly.pdf.