Journal of the American Medical Informatics Association, 25(3), 2018, 248–258
doi: 10.1093/jamia/ocx095
Advance Access Publication Date: 8 September 2017
Research and Applications

Research and Applications

A conceptual framework for evaluating data suitability for observational studies

Ning Shang, Chunhua Weng, and George Hripcsak

Department of Biomedical Informatics, Columbia University, New York, NY, USA

Received 31 March 2017; Revised 20 June 2017; Accepted 15 August 2017

ABSTRACT

Objective: To contribute a conceptual framework for evaluating data suitability to satisfy the research needs of observational studies.

Materials and Methods: Suitability considerations were derived from a systematic literature review on researchers' common data needs in observational studies and a scoping review on frequent clinical database design considerations, and were harmonized to construct a suitability conceptual framework using a bottom-up approach. The relationships among the suitability categories are explored from the perspective of 4 facets of data: intrinsic, contextual, representational, and accessible. A web-based national survey of domain experts was conducted to validate the framework.

Results: Data suitability for observational studies hinges on the following key categories: Explicitness of Policy and Data Governance, Relevance, Availability of Descriptive Metadata and Provenance Documentation, Usability, and Quality. We describe 16 measures and 33 sub-measures. The survey uncovered the relevance of all categories, with a 5-point Likert importance score of 3.9 ± 1.0 for Explicitness of Policy and Data Governance, 4.1 ± 1.0 for Relevance, 3.9 ± 0.9 for Availability of Descriptive Metadata and Provenance Documentation, 4.2 ± 1.0 for Usability, and 4.0 ± 0.9 for Quality.

Conclusions: The suitability framework evaluates a clinical data source's fitness for research use. Its construction reflects both researchers' points of view and data custodians' design features. The feedback from domain experts rated Usability, Relevance, and Quality categories as the most important considerations.

Key words: data suitability, survey, observational studies.

BACKGROUND AND SIGNIFICANCE

Wide adoption of electronic health records (EHRs) has contributed to the amassing of large amounts of patient data.[1] Along with the emerging open data policies, data are increasingly available for research.[2] Clinical data repositories can provide enormous benefits to a health care organization in management and strategic decision-making, patient-specific clinical decision support,[3,4] and evidence-based research improving point of care. This promises to accelerate medical discovery through observational studies.[5] For privacy protection, researchers need to determine whether the dataset is appropriate for the intended research before obtaining institutional review board (IRB) approval to access the data. This is a time-consuming process, and the suitability of using data for research of interest cannot be guaranteed. In this context, suitability is defined as the fitness of a clinical dataset for the intended purpose, specifically the extent to which the dataset can meet research needs for observational studies. Despite comprehensive research reuse of clinical databases, discussion concerning data suitability is scarce in the biomedical informatics literature.[6] Current studies of clinical data primarily focus on data quality.[7–10] Data quality can be considered as a broad field that identifies data characteristics that are important to both data custodians and data users, using "fitness for use" as guidance. Data quality can also focus on one dimension of data, for example accuracy. Studies have shown that the quality of EHR data is highly variable.[10] Poor data quality can cause unreliable research results. However, data quality is not the sole determinant of data

© The Author 2017. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

effectiveness for research reuse. Data suitability focuses on data characteristics that can assist in summarizing researchers' data needs and mapping the research needs to the data provided.

To standardize suitability assessment, this study contributes an original conceptual framework that applies to observational studies.[6] It can be used by researchers to detect the potential unsuitability of a dataset before making expensive investments in data access. Using patient health data for research poses privacy risks and legal and policy challenges associated with sharing such data for research reuse.[11–13] To reduce the risks, our data suitability framework focuses on data characteristics that can be reported on an aggregate level using summarized data or on relevant metadata. In this way, the framework enables sharing of data suitability metrics with other researchers. This can help avoid unnecessary administrative processes by not having to get IRB approval before assessing data suitability. By the same token, our data suitability framework enables evaluation and comparison of heterogeneous data sources.

METHODS

The responsibilities of custodians of clinical data repositories are primarily to store and integrate data generated at the point of care and collected from EHR systems.[4] During these processes, knowledge on the data is accumulated. This knowledge can be expressed through data characteristics, properties possessed by or derived from a clinical database. Data characteristics in turn provide opportunities for data users (researchers) to understand data if they can be used for research.[14] The data users and the context of the usage determine the selection of data characteristics.

To understand the central needs of data custodians and data users, we conducted a literature review and a scoping review to address the 3 questions below:

- What tasks and resources are needed to access the data (ie, can the data be accessed)?
- What analyses do the data permit (ie, what is the content)?
- Are the data ready for research use (are the data usable)?

The overall methodology of developing the framework is illustrated in Figure 1.

Literature review of researchers' data needs

The first review was a systematic review of population-based observational studies using clinical databases (Figure 2). Our aim was to identify researchers' data needs and analyze clinical database characteristics pertinent to observational studies. In order to find relevant publications, a search was performed in PubMed in December 2014 using the following query: "(population-based observational study) AND ((electronic health records) OR (clinical data repository) OR (clinical data warehouse) OR database OR (electronic database) OR (large scale data sources))." We refined the search terms over multiple iterations, repeatedly reviewing the list of results with the aim to maximize the number of publications that reported on large-scale observational studies based on electronic clinical data. It should be noted that PubMed's search engine parses queries and maps terms and phrases to concepts, and it employs those concepts to build an enhanced query.[15] In our case, mapping to Medical Subject Headings terms and publication types helped to improve query results. For example, the query string "observational studies" was converted to a disjunctive expression including observational studies as a Medical Subject Headings term, a publication type, and a keyword for text search. A total of 218 papers were retrieved, of which only 117 papers that met the following criteria were included in the final analysis:

a. The study is an observational study,
b. clinical database is the main data source,
c. full text is available in English, and
d. data source–related issues were discussed.

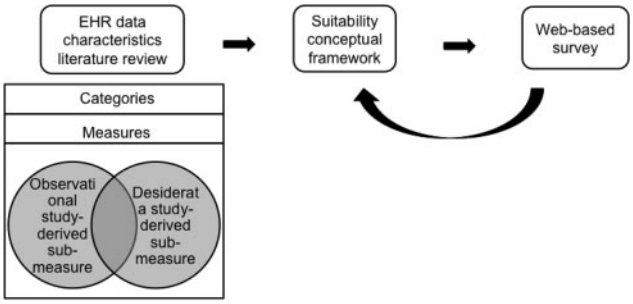

Figure 1. The workflow for developing the suitability conceptual framework. In constructing the framework, we conducted a literature review, and its findings were summarized and discussed within our team. A web-based survey from domain experts was used to validate the framework.

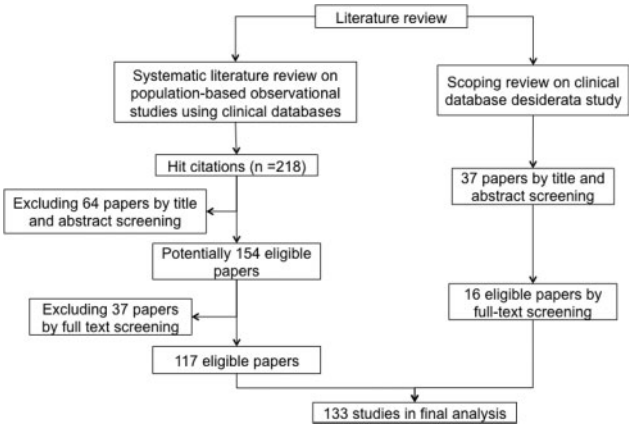

Figure 2. Paper selection process in literature review.

Scoping review of the desiderata for observational databases

The second review (Figure 2) was a scoping review – an iterative process to search for relevant manuscripts in a relatively new domain wherein subject headings and keywords are not yet well established[16,17] – of studies investigating the desiderata for clinical databases that deal with the information storage model, terminology, data integration, and value set management to meet different data use requirements of researchers, clinicians, and administrators.[18] The review aimed to understand the design ideas for clinical data repositories and potential gaps when using the repository for research. We queried "clinical data warehouse characteristics" and "clinical data warehouse assessment for research reuse" in Google Scholar and extended the results by reviewing the citations in the identified articles. A total of 37 articles were retrieved, 16 of which were relevant.

Construction of the suitability framework

Each review revealed a list of operational features that can be used to assess the suitability of clinical databases for observational studies. We refer to operational features as sub-measures, which are fine-grained data characteristics that sometimes show overlap or ambiguities and hence are useful for harmonization to define suitability measures. We grouped them into categories. A category is a conceptual grouping of entities that are considered similar by relating to the same facet of the data. In iterative discussions, we vetted and grouped the identified sub-measures. This bottom-up approach facilitated recognition of the data needs of actual data users, built the framework from known components that were published from reliable sources, and directly linked the conceptual framework to a concrete metric. The latter can form a basis for assessing scores in these categories.

Suitability for a purpose depends on the characteristics, or qualities, of the data. In a well-known data quality framework, Wang et al.[8] observed the multifacetedness of data characteristics and summarized them into intrinsic, contextual, representational, and accessible. During the framework development, we used these 4 facets as guidance to explore the relationships between the identified suitability categories.

Evaluation of the suitability framework

To evaluate the framework for the importance of the individual components that it comprises, an anonymous IRB-approved web-based survey was conducted. Subjects were recruited via the following e-mail lists: Observational Health Data Sciences and Informatics, American Medical Informatics Association Clinical Research Informatics, and Biomedical and Healthcare Data Discovery and Indexing Ecosystem. To qualify answers for inclusion in the final survey analysis, 5 background questions determined the subject's amount and type of experience in exploring and using clinical data for biomedical research studies, or in providing or managing clinical data. The survey was open for 8 weeks from July to September 2015, and final reminders were sent at the beginning of the sixth week. The use of e-mail lists made it impossible to determine the response rate.

The main survey instrument comprised attitude questions about the framework categories and sub-measures (referred to as attributes and presented in a simplified way for easy comprehensibility). The attitude questions in Likert-type scales[19] required respondents to indicate whether they had positive or negative impression about the framework's components.

To pretest the questionnaire, 7 fellow researchers with data use or survey design experience were asked to provide feedback on all aspects in order to identify errors and areas for improvement. In addition to improvements in wording, the pilot study also led to moving the suitability category from the beginning to the end. Since an understanding of the framework develops and improves while progressing through the survey, the insights that were gained support giving more informed and fair feedback. Most important of all, the pilot study also evaluated the Likert-type scale item statements. Since all items were evaluated as important, we revised the scale to have 4 levels of importance and 1 level of unimportance instead of 2 levels each of importance and unimportance and 1 neutral.

The survey results were analyzed by calculating the mean importance ratings for the categories. A diverging stacked bar chart[20,21] was used to visualize the responses for different sub-measures. To find the most favorably rated suitability sub-measures, we calculated top 2 box scores. We also described respondents' experiences and their role in using EHR data for research.

RESULTS

This section is structured as follows. The subsection "Suitability considerations identified in the literature" summarizes our findings on the suitability considerations from the literature reviews. These findings were used to formulate the foundations for the suitability framework. The subsection "The suitability conceptual framework" presents the framework, and "Survey evaluation results" concludes by presenting the evaluation results.

Suitability considerations identified in the literature

Literature review on researchers' common data needs in observation studies

We found that to demonstrate the usefulness of data for a study, researchers often cite other scientific studies that were performed on the same database, especially if the research interest (eg,[22]) is related. Similarly, to illustrate generalizability, researchers cite other studies that describe the demographic distribution of the same database. The intent is to show that the data can represent the general or a research-specific population (eg,[23]), and thereby to decrease the data's effect on selection bias.[24] An example for a well-established database with archived evidence of database representativeness and usefulness is the General Practice Research Database.[25]

Since an EHR is designed for clinical practice, its data often do not include all the variables required for observational studies, which can potentially cause misclassification bias.[26] Consequently, if this is of concern to researchers, they also report which variables of interest are not available (eg,[27,28]) to illustrate the potential bias of the data.

Temporal factors are reported to be essential for research methodologies at the operational level,[29] not only for retrieving and measuring study variables, but also as having an impact on data characteristics. For example, laparoscopic procedure codes have been available only since 2003, which explains the study cohort selection time constraints in[30].

Researchers using routinely collected data frequently limit their study reproducibility based on factors such as the quality of available data.[31] Sometimes, differences between patients with complete and incomplete data across exposure and outcome variables are reported (eg,[28]). Also, published papers are cited to calibrate data accuracy for their studies (eg,[25,32–34]). Data completeness and accuracy are key considerations when using clinical databases for research.

Recent studies increasingly emphasize data provenance, especially for curated databases.[35] It is reported that a lack of this consideration could result in misunderstandings of patient information out of context.[36] Data provenance is usually described as what information is contained, where it comes from (eg, which hospital, department, specialty, visit, EHR system[23,37]), when the database is established (eg,[25]), and how the data are entered and by whom (eg,[28]). Qualitative research methods, such as reviews of clinical documentation, interviews with data custodians, or direct observations of workflow, can be used to retrieve the relevant information.[38] For studies involving multiple databases, data linkage processes are also described (eg,[22,39]).

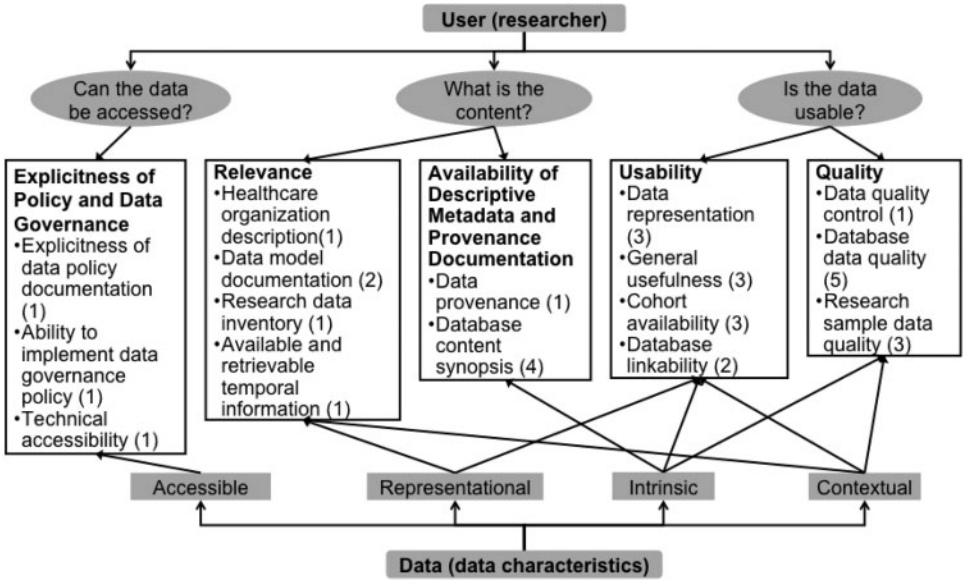

Figure 3. The suitability conceptual framework (categories and measures) for evaluating the suitability of electronic clinical data for observational studies.

Scoping review on frequent clinical database design considerations

In addition to data users' research needs for clinical databases, we also identified the design features employed by data custodians that have effects on data suitability.

The desiderata study revealed different methods to enable access to clinical data. In[40] these are based on whether the information desired is from aggregate, deidentified, or fully identifiable patient data. It is recommended that all aspects and steps of data extraction should be planned based on a data use protocol and clearly stated in a specification document.[41] Hu et al.[42] and Shin et al.[43] also considered constructing data access or direct query interfaces to view and analyze data. In addition, an up-to-date link to the custodian who is most knowledgeable about the data repository should be provided.[18] The above considerations are intended to solve technical issues in data accessing. To comply with data use policy, protect patient privacy, and guarantee data security, it is often recommended that related regulatory considerations and ethical principles should be well documented.[18,41,42,44]

When data are collected across different EHRs or multiple institutions,[18] describing the original context and data provenance is considered important for secondary use.[45] In order to integrate new data sources into existing schemas,[18,46] data custodians document the source of all crucial fact tables and data extract, transform, and load processes to help understand data origins and enable choice of the right data subsets.[18,46] Furthermore, an understanding of the longitudinality of patient records and the time between the occurrence of clinical events and their availability for analysis should be provided,[41] All these aspects are reported as usually being affected by factors such as repository size, annual growth, and data collection.[3,40,41] Since different variables are required for different studies, summarizing the types of data contained in a clinical data repository can assist in determining whether a variable of interest is captured and can later be extracted.[41,47]

Many research studies require the use of multiple resources to identify an adequate study cohort or to formulate a complete patient story.[41,48] In such scenarios, database-level and person-level record linkages are important, and data custodians should provide means to facilitate such linkages.

As the main layers of clinical data warehouse architecture, terminologies standardize the representation of medical concepts,[18,42] and data models are carefully designed to store clinical data. To make the data easily used for research, data custodians are required to document vocabularies used, make the data dictionaries available,[45] and do similarly for data models.[18] The Observational Medical Outcomes Partnership Common Data Model[49] has emerged as a popular model for assisting with data standardization for research.

To decrease information loss and data errors, the data processing procedures of clinical data warehouses usually contain data quality assurance components.[40,45]

The suitability conceptual framework

The suitability framework (Figure 3, with details in Table 1) encompasses 5 categories to address the 3 key questions to satisfy researchers' data needs. Explicitness of Policy and Data Governance is an important determinant of suitability for research. Relevance and Availability of Descriptive Metadata and Provenance Documentation inform researchers what they are allowed to do with the data. Usability and Quality assure that data can be used.

The framework is divided into categories, measures, and submeasures. It also establishes the relationships among categories using the data facets represented in Wang et al.'s[8] framework. Suitability categories can be related to one another, since each one can reflect one or more facets of data characteristics, and a facet can be discussed in one or multiple categories. In the following, the 5 categories are explained in detail, followed by definitions of the data facets and an explanation of how they relate the categories to each other.

Explicitness of policy and data governance

In a general sense, data policy addresses issues related to data access and data governance. This often includes nontechnical aspects like ownership and user agreements as well as technical aspects like permission control and security. Data policies can serve as broad guidelines for governing data and can also be used to develop strategies for data access. They can help answer questions such as whether

Table 1. The suitability conceptual framework

Category (definition)	Measure (definition)	Sub-measure (definition)	Example
Explicitness of Policy and Data Governance: deliberate principles to guide users (researchers) in accessing clinical data	Explicitness of data policy documentation: relative ease in obtaining data governance policy	Data access, privacy, and security policy: describes the authorization policy and privileges each user has for accessing the data	Documented data access policy is available for users
	Ability to implement data governance policy: relative ease in implementing data governance policy	Data access, privacy, and security system: availability of systems that can protect privacy and guarantee security	Data custodians provide a secure platform for users to access clinical databases[43]
	Technical accessibility: data are technically accessible and available for retrieval	Technical accessibility: relating to querying, extracting, transferring, and storing data	Data can be extracted for research[50]
Relevance: to evaluate whether the data and pertinent information contained are appropriate for research of interest	Health care organization description: descriptive data about health care organization	Health care organization descriptive data: summary of health care organization–related data	Health care organization (facility, provider, payer) overview is provided to users[18]
	Data model documentation: understand how data are modeled into what types for further retrieval	Data types included: specification of data types included in the data warehouse	Patients' textual clinical notes and histopathological imaging are included in the clinical data warehouse[51]
	Research data inventory: an inventory of available core data elements of clinical database for research	Capturing common data elements: describes the availability of study variables (external knowledge) that are pertinent to clinical research	The existence of general research data inventory (data groups/data elements, eg, EHR4CR data inventory) in clinical databases is specified[32]
	Available and retrievable temporal information: the age of the data and temporal information contained are obtainable	Data historical evolution: describes historical evolution of data capture and related changes	First-time procedure captured in clinical databases and time trend of procedure use in clinical databases have been recorded[30,47]
		Retrievable temporal information: temporal information is available and retrievable for variables of interest	Existing time stamps of research variables of interest can be used to retrieve the variables for defined study period[52]
Availability of Descriptive Metadata and Provenance Documentation: documented metadata to identify a resource and describe its intellectual content and information origin	Data provenance: refers to attributes of the origin of health information, facilitates tracking relevant events of data transformation from the original state	Database origin and derivation: describes attributes of the origin of health information and its derivation history	The entities (eg, hospital, department) and processes (eg, extract, transform, load strategy) involved in producing data are well documented[28] Review of clinical documentation, interviews with data custodians, or direct observation of workflow can be used to gather data provenance[38]
	Database content synopsis: describes the content of a clinical database for the purpose of identifying appropriate database for clinical research	Database longitudinality: describes longitudinal records duration	The duration of clinical databases is recorded[41]
		Summary of the database: specify database size, architecture, and update cycles	Clinical database size and data update cycles are recorded[3]
		Data about clinical system: specify original health data collection systems	The clinical systems that feed the clinical databases are specified[40]
		Data components summary: describes data components (referring to diagnosis, for example) and their environment	Clinical databases' stored data groups (eg, diagnosis, procedure, drug, lab) should be specified[18]
Usability: data are able to be used for research of interest	Data representation: illustrate how data and information from real-world clinical practice are mapped, represented, and stored in the database	Data storage model: describes how different types of data are stored	Data storage format (eg, coded field, free text, scanned document, links to a picture archiving and communication system) for data objects in clinical databases are specified[48]
		Data standards: specify existing common data standards	Clinical databases incorporating standard data model should be specified[40]
		Terminology/semantic interpretation: specify existing terminology layer	Clinical databases terminology layer (representing coded medical concepts) should be specified[45]
		Standardized coding for variables of interest: available coded data for variables of interest	Standardized coding can be used to retrieve data, to ascertain research-related variables (eg, outcomes, covariates)[28]

(continued)

Table 1. continued

Category (definition)	Measure (definition)	Sub-measure (definition)	Example
	General usefulness: the extent to which data are trusted in their utility for research	Database research reuse by published literature: demonstrate the capability of database reuse for research by archiving studies that were conducted with the database	Published studies show that clinical databases are appropriate for research of interest[23]
		Representative general population by published studies or external standard: the database can represent general population that is supported by published studies or by comparing with external standard	Cited reference is used to illustrate that the geographic, age, and gender distributions of the database population represent the general population[25]
		Available research required subpopulation: required subsets of the general population for research are contained in the database	A database that excludes those who are too sick to work may not be suitable for studying renal failure medications[41]
	Cohort availability: the volume of available data is appropriate for research of interest	Sufficient study sample size: enough samples can be retrieved to conduct the research	The number of available subjects from a database is sufficiently large to meet research objectives[41]
		Available required variables of interest: capture study variables that are essential to the research of interest	Researchers can identify all variables hypothesized to influence the outcome of interest from clinical databases[47]
	Database linkability: describes whether a database can link to original data source or other existing databases and how this can be achieved	Person-level linkability: person-centered data can be identified across the health organization	Clinical databases can link disparate databases (registries, clinical databases in different organizations) for patients[41]
		Database-level linkability: describes comparability or compatibility among multiple databases for research use	Disparate clinical databases (eg, registry, claims, EHR data) are compatible to link with each other[44]
Quality: an essential characteristic that determines the reliability of data for research	Data quality control: there exists a component of assuring data quality in constructing the database	Documented data quality assurance components in data warehouse extract, transform, load processes	Data quality assurance component exists in data derivative process and data quality control results are documented[53]
	Database data quality: the reliability of the overall database	Database data quality by published data quality study: provide literature evidence to prove the data quality of the database	Cite previous literature in which the database's reliability and validity have been examined[47]
		Records completeness: defined through documentation, breadth, density, and predictive completeness[54]	Distribution of clinical data points is demonstrated to be complete by evaluating documentation, breadth, density, and predictive completeness[54]
		Data accuracy by external validation: validate database records against external standards	Discrepancy of prevalence calculated from clinical databases and national or published prevalence data might indicate incomplete recording[41,55]
		Data accuracy by internal validation: validate database records utilizing interdependent relationships among data points	Disease prevalence calculated from diagnosis code and approximate codes (eg, measurement indicative of disease) is similar[41,56]
		Data accuracy by logical checking: validate database records using logic rules	Data element has a value that is different from the norm by predefined logic rules (eg, age- or gender-specific disease cannot be recorded for non-indicated age or gender group)[41]
	Research sample data quality: the reliability of sampled data that are of interest for research	Systematically captured variables: systematic capture of variables was not biased	Clinical databases can record abnormal test results but not always normal findings[41]
		Cohort variable quality: validated quality for study-specific variables and selected cohort	Data-quality checks are conducted for key variables used for study population (cohort) identification[45]
		Coding granularity: level of detail or specificity used to encode concepts is appropriate for research of interest	Granularity of coding (eg, type of diabetes is specified) is appropriate and accurate for study variables[41,48]

data marts need to be developed or data can be accessed directly in an existing warehouse or another data source.[57–59]

The nature of data in the clinical domain calls for stricter and more elaborate policies to protect the privacy and security of individuals' information. Data policy specifically requires addressing issues related to data ownership and custodianship, data use and liability (end-user agreements and licenses), data access (permission control), data security, data acquisition (governed through IRBs), and patient consent. In addition, Health Insurance Portability and Accountability Act laws strictly regulate health information privacy.

It is important for data users to be aware of how data can be used and, more important, that data administration guided by data policies can ensure data access and use in a controlled fashion. The importance of discussing data access, use, and control is emphasized through data life cycles.[60]

Relevance

A big volume of data does not mean the data are relevant to the research of interest. Researchers need to be able to understand how to discern what information is relevant for their analyses. The suitability framework therefore provides views that aim to assist in this discernment.

The Relevance category specifies metadata needed to provide a database overview that enables choice of an appropriate database. Metadata in this sense not only refers to storage-related aspects like database schemas, but also includes descriptive information about the data. To evaluate whether the data and pertinent information they contain are appropriate, the framework provides measurements for 3 aspects: health care organization description, data modeling, and data content.

Clinical data capture in medical practice is affected by physicians, hospitals, systems, and workflows – what data are collected and how they are collected. For example, the patient population covered by different facilities may not be the same. Clinical data from a nursing home setting are less likely to be used to investigate disease prevalence in the pediatric population. Background information about the health organization will help researchers determine whether their target cohort and necessary clinical data can be provided.

Data organization describes the formats in which clinical data are stored (eg, structured, coded, textual, images, electrocardiogram signals). This knowledge is helpful for researchers to select tools or methods needed to retrieve the data. Understanding the data organization paves the way for correctly extracting relevant data.

The essence of relevance lies in the actual data content, which mainly consists of research data inventory and temporal information. The data inventory represents data elements available for clinical research and can help in evaluating a study's feasibility.[61] Recent EHR implementations have focused on building a longitudinal health record to provide a clinical history of patient-based clinical episodes over time. Longitudinal data provide crucial temporal information for evidence-based medical care and enable secondary use for clinical research. The importance is reflected in the fact that temporal constraints are present in 38% of clinical research eligibility criteria.[62]

Alongside data, documenting historical evolution information (eg, implementation of new data collection systems) is important to avoid artificial data patterns in data analyses.[18]

Availability of descriptive metadata and provenance documentation

Adequate information is indispensable for researchers to correctly interpret data. This emphasizes the importance of documentation of metadata and data provenance.[63] Metadata is "data about data"; specifically, it is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.[64] Descriptive, structural, and administrative metadata are 3 main metadata types.[64] Structural and administrative metadata provide information to manage resources, while descriptive metadata enables identification and characterization of resources. The latter is more relevant to researchers.[64]

To identify appropriate databases for observational studies, descriptive metadata is critical. Several relevant measures were identified during the literature review[23,37] and are referred to as database content synopsis in this framework.

Data provenance[65,66] pertains to the derivation history of a data product starting from its original sources[65] and helps researchers understand the authoritative source(s) of a given variable of interest.[25,48] Both database content synopsis and data provenance are metadata, so we put these 2 measures into this category.

Usability

The existence of collected health data does not necessarily mean that the data can be used for research, which can depend on the user or the specific research topic.[67] Measures of data representation, specifically data storage, standards, and terminology, are considered important aspects of usability.[18,40,41]

Knowledge about data usefulness for a specific research area or population can help data users determine the appropriateness of a database for their research.[22,25,41] This kind of knowledge about data can be gathered from existing publications and accumulated by data custodians to show to potential users. The general usefulness of data is directly related to what research the data can be applied for.

A precise estimation of the association between exposures and outcomes in observational studies is usually expressed as a confidence interval, which is determined by the sample size.[68] This emphasizes the importance of the availability of sufficient numbers of cohorts for observational studies.[41,69]

In reality, patients' longitudinal medical encounters occur across multiple settings or institutions and involve multiple care providers. Restructuring the data through linkable electronic clinical data is important to understand pathways of care and to allow researchers to investigate and conduct experiments in an efficient and valid manner.[39,41,44,70]

Quality

Data quality is a research question of its own. Juran introduced "fitness for use" as a universal concept for quality as "the extent to which a product successfully serves the purposes of the user."[7,71] This definition is widely applied in different fields[72] and adopted frequently when developing data quality frameworks.[8,10]

In our framework, data quality is restricted to the scope of researchers using clinical databases for observational studies, with a focus on the quality of a clinical database for specific research questions. For example, a database suitable for observational studies of diabetes may not suit predictive modeling of diabetes progression. Reviewing population-based observational studies that use clinical databases, we noted that researchers use published evidence to describe the accuracy and completeness of databases and study variables to reflect quality. The desiderata review emphasizes the importance of (1) the data quality control component in data derivative processes, (2) accuracy by internal, external, or logical validation, and (3) systematically captured variables and their coding. Consequently, quality assessment in the suitability framework was

defined by the quality assurance component, database completeness, database accuracy, and quality of study population and study variables sub-measures. These sub-measures are grouped into data quality control, database data quality, and research sample data quality.

Relatedness of the categories

To represent general data characteristics, we adjusted the 4 aspects describing data multifacetedness from Wang's framework[8]:

1. Intrinsic: describe the objective existence of data.
2. Contextual: describe data within the context of the task at hand.
3. Representational: describe the format and model.
4. Accessible: describe through what procedure those data can be accessed to users.

Explicitness of Policy and Data Governance represents access. Relevance is determined by both appropriate clinical data and retrievable data representation to decide whether clinical data are suitable for research of interest. Availability of Descriptive Metadata and Provenance Documentation objectively describes data origins and summaries of existing data. Usability and Quality depend on researchers' interest, so both consider whether context is appropriate for the task; at the same time, they both depend on whether the data are intrinsically good. In

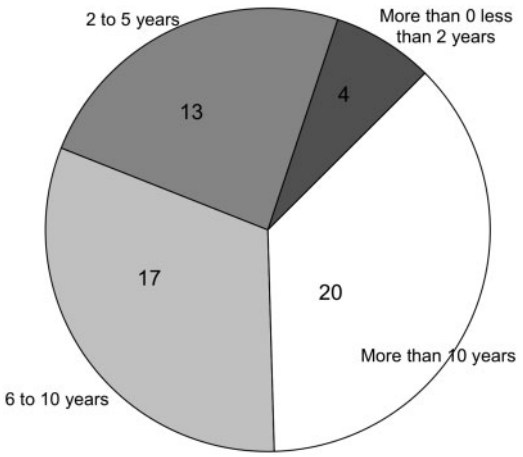

Figure 4. Respondents' experience of data use.

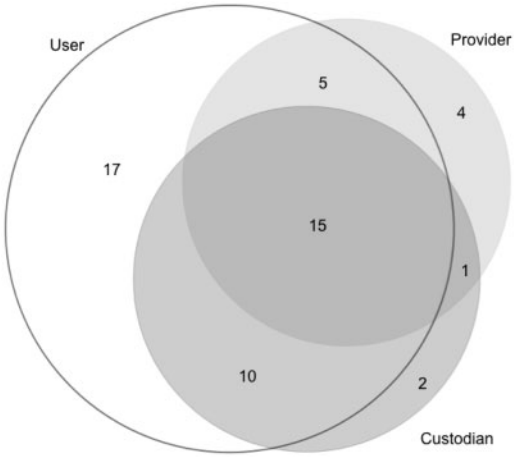

Figure 5. Multiple roles of data users/providers/custodians among survey respondents.

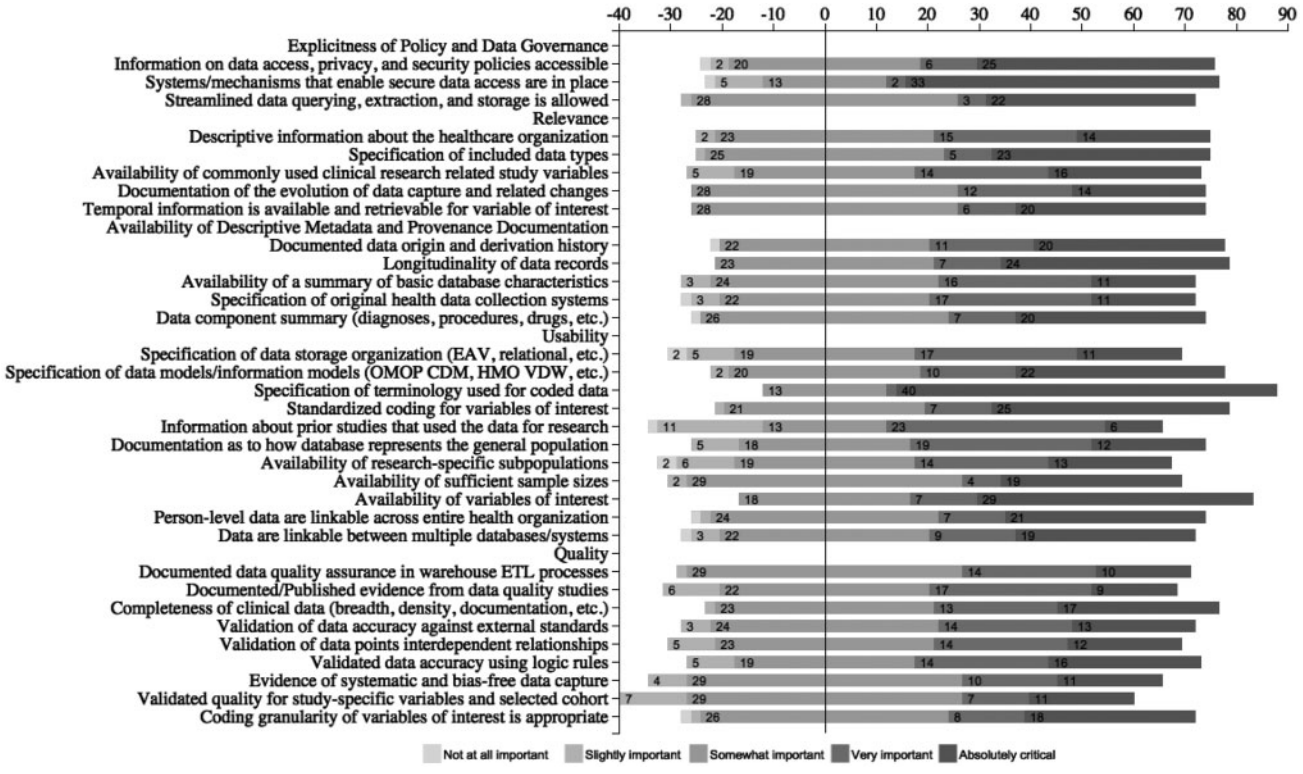

Figure 6. Web-based survey results for evaluating the importance of the suitability framework sub-measures.

addition, data representation is important for data being a good use, so Usability also reflects the representational characteristic.

Survey evaluation results

As explained earlier, the survey evaluated the framework for the importance of categories and sub-measures by experienced data users. The survey was completed by 54 respondents, of whom 37 had >5 years of data usage experience (Figure 4), and 47 were data users (Figure 5).

The resulting importance ratings of categories are as follows:

- 3.9 ± 1.0 for Explicitness of Policy and Data Governance
- 4.1 ± 1.0 for Relevance
- 3.9 ± 0.9 for Availability of Descriptive Metadata and Provenance Documentation
- 4.2 ± 1.0 for Usability
- 4.0 ± 0.9 for Quality

Across the 33 sub-measures, the rating frequency of very important or absolutely critical ranged from 33% to 76% (Figure 6). Based on this top 2 box score, the top 5 most rated suitability sub-measures were:

- Specification of terminology used for coded data (76%)
- Availability of variables of interest (67%)
- Systems and mechanisms that enable secure data access are in place (65%)
- Standardized coding for variables of interest (59%)
- Specification of data models/information models (59%)

DISCUSSION AND CONCLUSIONS

The suitability conceptual framework captures key considerations for choosing suitable clinical databases to conduct observational studies. It is built upon a literature review of researchers' data needs using clinical databases and a scoping review of desiderata for developing clinical databases, and is validated by a survey of domain experts. The framework reflects the perspectives not only of data users conducting observational studies, but also of data custodians building clinical databases with research reuse in mind. In addition to validating the framework, experts' opinions were elicited on the importance of categories and sub-measures. They rated Usability, Relevance, and Quality as the most important categories in this framework.

Our framework can assist researchers in determining whether a database is likely to be suitable for a selected research question. It aims to serve as a critical component for a formal assessment of databases for observational studies and a potential guideline to build a suitability index for quantifying and comparing the suitability of clinical databases for use in research topics.

FUNDING

This work was supported by a Data Discovery Index Coordination Consortium Administrative supplement award (3R01LM006910-15S1) to parent grant "Discovering and Applying Knowledge in Clinical Databases" (PI: Hripcsak, R01 LM006910).

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

All authors contributed to the conception of the study. GH is the grant holder. NS conducted the literature review and implemented the web-based survey. All authors contributed to refinement of the study protocol, development of the conceptual framework, and design of the web-based survey. All authors contributed text to the paper and approved the final manuscript.

ACKNOWLEDGMENTS

We thank the survey respondents recruited from the Observational Health Data Sciences and Informatics, American Medical Informatics Association Clinical Research Informatics, and Biomedical and Healthcare Data Discovery and Indexing Ecosystem communities.

REFERENCES

1. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med*. 2010;363:501–04.
2. Margolis R, Derr L, Dunn M, *et al*. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*. 2014;21:957–58.
3. Schubart JR, Einbinder JS. Evaluation of a data warehouse in an academic health sciences center. *Int J Med Inf*. 2000;60:319–33.
4. Evans RS, Lloyd JF, Pierce LA. Clinical Use of an Enterprise Data Warehouse. *AMIA Annu Symp Proc*. 2012;2012:189–98.
5. Thiese MS. Observational and interventional study design types: an overview. *Biochem Medica*. 2014;24:199–210.
6. Sørensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol*. 1996;25:435–42.
7. Juran JM. *Juran's Quality Control Handbook*. 4th ed. New York: McGraw-Hill; 1988.
8. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. *J Manag Inf Syst*. 1996;12:5–33.
9. Kahn MG, Raebel MA, Glanz JM, *et al*. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care*. 2012;50. www.ncbi.nlm.nih.gov/pmc/articles/PMC3833692/. Accessed October 24, 2014.
10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20:144–51.
11. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13:395–405.
12. Hripcsak G, Bloomrosen M, FlatelyBrennan P, *et al*. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. *J Am Med Inform Assoc*. 2014;21:204–11.
13. 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.
14. Vullings W, Meijer M, Bulens J, *et al*. Spatial Data Quality: What do you mean? In: *AGILE Conference Paper*. Lisbon; 2015: 9–12.
15. *How PubMed Works: Automatic Term Mapping*. National Center for Biotechnology Information (US); 2017. www.ncbi.nlm.nih.gov/books/NBK3827/. Accessed August 9, 2017.
16. Embi PJ. Clinical research informatics: survey of recent advances and trends in a maturing field. *IMIA Yearb*. 2013;8:178–84.
17. Levac D, Colquhoun H, O'Brien KK, *et al*. Scoping studies: advancing the methodology. *Implement Sci*. 2010;5:1–9.
18. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. *AMIA Annu Symp Proc*. 2013;2013:648–56.
19. Desselle SP. Construction, implementation, and analysis of summated rating attitude scales. *Am J Pharm Educ*. 2005;69:1–11.
20. Heiberger RM, Robbins NB. Design of diverging stacked bar charts for Likert scales and other applications. *J Stat Softw*. 2014;57:1–32.

21. Robbins NB, Heiberger RM. Plotting Likert and other rating scales. In: *Proc 2011 Joint Statistical Meeting*. 2011. 1058–66. www.amstat.org/sections/SRMS/proceedings/y2011/Files/300784_64164.pdf. Accessed January 12, 2017.
22. Visser ST, Schuiling-Veninga CC, Bos JH, *et al*. The population-based prescription database IADB.nl: its development, usefulness in outcomes research and challenges. *Expert Rev Pharmacoecon Outcomes Res*. 2013;13:285–92.
23. Lee M-TG, Lee S-H, Chang S-S, *et al*. Comparative effectiveness of different oral antibiotics regimens for treatment of urinary tract infection in outpatients: an analysis of National Representative Claims Database. *Medicine (Baltimore)*. 2014;93:e304.
24. McVeigh KH, Newton-Dame R, Perlman S, *et al*. Developing an electronic health record-based population health surveillance system. *NY City Dep Health Ment Hyg*; 2013.
25. Amar RK, Jick SS, Rosenberg D, *et al*. Incidence of the pneumoconioses in the United Kingdom general population between 1997 and 2008. *Respiration*. 2012;84:200–06.
26. Haneuse S, Daniels M. A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs* 2016;4:1203.
27. Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prev Sci Off J Soc Prev Res*. 2013;14:570–80.
28. Tomasson G, Peloquin C, Mohammad A, *et al*. Risk for cardiovascular disease early and late after a diagnosis of giant-cell arteritis: a cohort study. *Ann Intern Med*. 2014;160:73–80.
29. Kelly J, McGrath J. *On Time and Method*. Newbury Park, CA: SAGE Publications; 1988. http://methods.sagepub.com/book/on-time-and-method. Accessed January 23, 2017.
30. Vricella GJ, Finelli A, Alibhai SMH, *et al*. The true risk of blood transfusion after nephrectomy for renal masses: a population-based study. *BJU Int*. 2013;111:1294–300.
31. Benchimol EI, Smeeth L, Guttmann A, *et al*. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med*. 2015;12:e1001885.
32. Chao P, Shih C-J, Lee Y-J, *et al*. Association of postdischarge rehabilitation with mortality in intensive care unit survivors of sepsis. *Am J Respir Crit Care Med*. 2014;190:1003–11.
33. Sant M, Minicozzi P, Mounier M, *et al*. Survival for haematological malignancies in Europe between 1997 and 2008 by region and age: results of EUROCARE-5, a population-based study. *Lancet Oncol*. 2014;15:931–42.
34. Fleet JL, Weir MA, McArthur E, *et al*. Kidney function and population-based outcomes of initiating oral atenolol versus metoprolol tartrate in older adults. *Am J Kidney Dis*. 2014;64:883–91.
35. Buneman P, Khanna S, Wang-Chiew T. Why and where: A characterization of data provenance. In: *International Conference on Database Theory*. Springer; 2001: 316–30. http://link.springer.com/chapter/10.1007/3-540-44503-X_20. Accessed December 28, 2016.
36. Johnson KE, Kamineni A, Fuller S, *et al*. How the provenance of electronic health record data matters for research: a case example using system mapping. *EGEMS*. 2014;2:1058.
37. Mansi I, Frei CR, Pugh MJ, *et al*. Psychologic disorders and statin use: a propensity score-matched analysis. *Pharmacother J Hum Pharmacol Drug Ther*. 2013;33:615–26.
38. Smerek MM. *Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0)*. www.nihcollaboratory.org/Products/Assessing-data-quality_V1%200.pdf. Accessed February 6, 2015.
39. Kingwell E, Evans C, Zhu F, *et al*. Assessment of cancer risk with interferon treatment for multiple sclerosis. *J Neurol Neurosurg Psychiatry*. 2014;85:1096–102.
40. MacKenzie SL, Wyatt MC, Schuff R, *et al*. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *J Am Med Inform Assoc*. 2012;19:e119–24.
41. Hall GC, Sauer B, Bourke A, *et al*. Guidelines for good database selection and use in pharmacoepidemiology research: good database conduct in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf*. 2012;21:1–10.
42. Hu H, Correll M, Kvecher L, *et al*. DW4TR: a data warehouse for translational research. *J Biomed Inform*. 2011;44:1004–19.
43. Shin S-Y, Kim WS, Lee J-H. Characteristics desired in clinical data warehouse for biomedical research. *Healthc Inform Res*. 2014;20:109–16.
44. Dokholyan RS, Muhlbaier LH, Falletta JM, *et al*. Regulatory and ethical considerations for linking clinical and administrative databases. *Am Heart J*. 2009;157:971–82.
45. Collaborative DQ, Collaboratives EDM. *DQC White Paper Draft 1: A Consensus-Based Data Quality Reporting Framework for Observational Healthcare Data*. http://repository.academyhealth.org/cgi/viewcontent.cgi?article=1001&context=dqc Accessed February 6, 2015.
46. Sittig DF, Hazlehurst BL, Brown J, *et al*. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data. *Med Care*. 2012;50:S49–59.
47. Motheral B, Brooks J, Clark MA, *et al*. A checklist for retrospective database studies: report of the ISPOR task force on retrospective databases. *Value Health*. 2003;6:90–97.
48. Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51:S30–37.
49. Hripcsak G, Duke JD, Shah NH, *et al*. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574–78.
50. Jagsi R, Bekelman JE, Chen A, *et al*. Considerations for observational research using large data sets in radiation oncology. *Int J Radiat Oncol*. 2014;90:11–24.
51. Phan JH, Quo CF, Cheng C, *et al*. Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. *Biomed Eng IEEE Rev In*. 2012;5:74–87.
52. Stessin AM, Sison C, Schwartz A, *et al*. Does adjuvant radiotherapy benefit patients with diffuse-type gastric cancer? Results from the Surveillance, Epidemiology, and End Results database: RT for Diffuse-Type Gastric Cancer. *Cancer*. 2014;120:3562–68.
53. Macedo AF, Douglas I, Smeeth L, *et al*. Statins and the risk of type 2 diabetes mellitus: cohort study using the UK clinical practice research datalink. *BMC Cardiovasc Disord*. 2014;14:1.
54. Weiskopf NG, Hripcsak G, Swaminathan S, *et al*. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46:830–36.
55. Pringle M, Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. *Br J Gen Pract*. 1995;45:537–41.
56. Horsfield P. Trends in data recording by general practice teams: an analysis of data extracted from clinical computer systems by the PRIMIS project. *Inform Prim Care*. 2002;10:227–34.
57. Martín E, Ballard G. Data management best practices and standards for biodiversity data applicable to bird monitoring data. US North Am Bird Conserv Initiat Monit Subcomm Online. 2010. www.prbo.org/refs/files/12058_Martin2010.pdf. Accessed April 10, 2015.
58. Burley TE, Peine JD. *NBII-SAIN Data Management Toolkit. US Geological Survey*. 2009. https://pubs.er.usgs.gov/publication/ofr20091170. Accessed April 29, 2015.
59. National Land and Water Resources Audit. *National Land and Water Resources Audit: 2002–2008: Achievements and Challenges*. Canberra, ACT: National Land and Water Resources Audit; 2008.
60. Safran C, Bloomrosen M, Hammond WE, *et al*. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14:1–9.
61. Doods J, Botteri F, Dugas M, *et al*. A European inventory of common electronic health record data elements for clinical trial feasibility. *Trials* 2014;15:18.
62. Tu SW, Peleg M, Carini S, *et al*. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform*. 2011;44:239–50.

63. Shoshani A, Rotem D, eds. *Scientific Data Management: Challenges, Technology, and Deployment*. 1st ed. Boca Raton, FL: CRC Press; 2009.
64. National Information Standards Organization (US). *Understanding Metadata*. Bethesda, MD: NISO Press; 2004.
65. Simmhan YL, Plale B, Gannon D. A survey of data provenance techniques. Computer Science Dept., Indiana University, Bloomington, IN. 2005;47405. ftp://cuts.soic.indiana.edu/pub/techreports/TR618.pdf. Accessed February 2, 2015.
66. Lusignan S de, Liaw S-T, Krause P, *et al*. Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. *IMIA Yearb*. 2011;6:112–20.
67. Usable and/or Useful Data? https://canvas.instructure.com/courses/883420/quizzes/1043784. Accessed June 29, 2015.
68. Jepsen P, Johnsen SP, Gillman MW, *et al*. Interpretation of observational studies. *Heart*. 2004;90:956–60.
69. Leue C, Buijs S, Strik J, *et al*. Observational evidence that urbanisation and neighbourhood deprivation are associated with escalation in chronic pharmacological pain treatment: a longitudinal population-based study in the Netherlands. *BMJ Open*. 2012;2:e000731.
70. Thompson CA, Kurian AW, Luft HS. Linking electronic health records to better understand breast cancer patient pathways within and between two health systems. *eGEMs* 2015;3:1127.
71. Reeves CA, Bednar DA. Defining quality: alternatives and implications. *Acad Manage Rev*. 1994;19:419.
72. Bisgaard S. Quality management and Juran's legacy. *Qual Reliab Eng Int*. 2007;23:665–77.