NIH Collaboratory Rethinking Clinical Trials®

Health Care Systems Research Collaboratory

Acquiring and Using Electronic Health Record Data

Table of Contents

Prepared by: NIH Collaboratory Electronic Health Records Core with contributions from Meredith Nahm Zozus, PhD; Rachel Richesson, PhD, MPH; W. Ed Hammond, PhD; Gregory E. Simon, MD, MPH

Version: November 2015

Introduction

This resource white paper was developed in 2015 to introduce clinical researchers to using electronic health record (EHR) data for research, which is fundamentally different from using prospectively collected data, as has historically been done in randomized controlled clinical trials. Several aspects of EHR data drive these important differences, including the lack of control over data definitions and data collection processes in healthcare facilities, procedures for access and permission to use the data, frequent dependence on record linkage, the need for computable definitions for cohorts and outcomes of interest, and the intricacies of demonstrating that data are of adequate quality to support research conclusions. Further, data sharing in the context of secondary use of EHR data also differs in important ways from sharing prospectively collected data. This chapter covers these essential aspects of secondary use of EHR data in clinical research.

What It Means To Be a Secondary User of Health Record Data

Data contained in electronic health records (EHRs) are widely viewed as a potential treasure trove for medical research [1], although for decades researchers have expressed concerns about the suitability of health record data for such uses [2–5]. Nonetheless, clinical studies based on EHR data are on the rise due to the increasing availability of such resources—a circumstance due in large part to incentives from the Centers for Medicare and Medicaid Services (CMS) that encourage "meaningful use" of EHRs, as well as emphasis by the National Institutes of Health (NIH) via the Clinical and Translational Science Award (CTSA) program on the availability of health system data for research.

Because the medical record has historically been viewed as the "traditional source for clinical information," it has "almost been unquestioned as a source of clinical information for non-clinical purposes" such as research [6]. In addition, while there has not been a clear statement or directive for the assessment and reporting of data quality as an integral component of research results, this expectation is now being articulated in funding solicitations [7], and calls for the inclusion of data quality reports with research results are being heard [8,9].

A recent review focusing on the suitability of EHR data for use in healthcare quality measurement concluded that "…issues related to data accuracy, completeness, and comparability must be addressed before routine EHR-based quality of care measurement can be done with confidence [10]." Similar remarks in other peer-reviewed literature show that the selection of appropriate measures of data quality depend on the source of data that is used [11–13]. These same challenges await the use of EHR data for clinical research and may prove even more significant in this context. For example, in correlative studies using 'omics'-based assays (such as genomics or proteomics) where related sets of biological molecules are comprehensively studied in limited numbers of patients, the issues related to data quality, completeness, and comparability will be paramount.

Researchers have sought to use data from health records for clinical studies since the very early days of clinical recordkeeping [14]. The use of EHR data collected during the course of clinical care for research purposes is often referred to as a *secondary use* of healthcare data—that is, the data were first collected as part of routine patient care and will be secondarily used for research. Prior to the existence of EHRs, such data had to be manually abstracted from clinical documentation in order to make secondary use possible. Such processes were typically laborious, time-consuming, and error-prone [15].

However, the widespread availability of EHRs is now enabling access to health record data on a much larger scale. A single electronic query can return data from hundreds of thousands of patients in a matter of minutes, making it possible for researchers to answer important clinical and scientific questions quickly and efficiently, and at a fraction of the resource cost that would have been required using older methods. But just as with data collection for prospective studies, researchers must be able to demonstrate that these data are of sufficient quality to support the conclusions drawn from them. The methods for doing this are different for secondary use of existing data than for prospectively collected data.

Data are always an incomplete representation of the things and events they describe, and as such may be appropriate for some uses but inadequate for others. For this reason, the suitability of the available data must be assessed for each potential secondary use. For example, the purpose and setting of the data collection have many effects, such as the determination of which data should be collected, the choice of measurement or observation methods, the meanings assigned to the data values, the amount and kind of contextual information (metadata) retained, the timing of data collection, and the level of detail (granularity).

A researcher engaging in secondary use of EHR data typically has no control over the original collection of the data, which may have occurred years earlier. In addition, the researcher may be one or more steps removed from the original data as collected in the healthcare setting. The further removed that the research team is from the original data collection (date or process), the greater potential there is for misunderstanding, degradation, and loss of information.

For example, claims data that have been dictated by a clinician in a discharge summary and subsequently coded with a standard terminology (e.g., International Classification of Diseases [ICD] or Current Procedural Terminology [CPT]) represent processed data that are removed from their origin. Such data are at risk for information loss through data reduction from coding, through disassociation with contextual information, or through the introduction of error. Thus, while users of secondary data may not have control over the original data collection, they should understand how and why those data were originally obtained, as well as any subsequent processing to which they were subjected. Understanding these aspects of the data will help the researcher determine whether EHR data are of suitable quality for a particular study.

There are a number of steps involved in appropriately using EHR data for research, including 1) gaining permission to use the data, 2) assessing the availability of data for a research need, 3) identifying the needed data for the population of interest, 4) linking data from different sources, 5) assessing the quality of the data, 6) managing the data for the duration of a given study, and 7) archiving or sharing data after a study. The rest of this chapter is organized according to these stages of EHR data use.

Gaining Permission to Use Healthcare Data

Most healthcare organizations have procedures in place that define the permissible internal uses of the data they collect and store. These procedures govern data access for members of the care team, information exchange for care transitions, data use in quality improvement (QI) projects, and administrative reporting for organizational management. For some healthcare facilities, these categories constitute routine data use. Larger facilities that conduct research, such as academic medical centers or facilities with embedded researchers, also have procedures in place for secondary use of health system data for research. Secondary use of such data is governed by federal regulations and by procedures established by the facility's institutional review board (IRB), although not all secondary uses of data require oversight or consent. For example, investigators who are initiating a research project are often allowed limited access to explore health system data in order to assess the feasibility of a proposed study and address questions. They might query data to determine whether there are sufficient numbers of patients with a given condition within the health system to support a study.

However, the actual use of patient data for research may fall under a different level of oversight than that which covers such preliminary assessments. In the case of research, there are additional requirements, such as informed consent on the part of the patient, de-identified data, or the use of a limited dataset that cannot be re-linked to individual patients. In the United States, using patient data for research requires IRB approval if the study is a clinical investigation that supports applications for research or marketing permits for products regulated by the Food and Drug Administration (FDA; 21 CFR Parts 50 and 56), or more broadly, research involving human subjects conducted, supported, or otherwise subject to regulation by any federal department or agency (45 CFR part 46 [the Common Rule]). In addition, the Federalwide Assurance (FWA) for the Protection of Human Subjects requires that at all institutions that receive federal research funding have oversight from an IRB. At these institutions, IRB approval is needed to access and use the data, and departmental or organizational approval may also be required.

Some research studies are conducted by institutions that do not fall into the categories described above. However, in order to publish research involving human subjects, virtually all peer-reviewed journals require that the study must have been reviewed and approved

by an IRB or ethics board [16]. Thus, investigators are advised to locate an appropriate IRB prior to embarking on research using patient data, even if their institution does not require it.

Additional contractual agreements and regulatory compliance are required when investigators want to use data from institutions that they are not directly associated with (e.g., a university researcher who wants to use data from local community hospitals). The Health Insurance Portability and Accountability Act (HIPAA) requires that *covered entities* and their business associates may release protected health information (PHI) only in certain controlled situations, including release to healthcare reimbursement departments or operations, to individual patients, to regulatory authorities, for national priority purposes, with authorization from the individual, and as a limited dataset. In these situations, the minimum necessary information may be released.

Covered entities include health plans, healthcare clearinghouses, and healthcare providers who electronically transmit any health information in connection with transactions for which the US department of Health and Human Services has adopted standards.

In the specific case of research, de-identified data sets may be used to share data in many cases, and, in addition to the regulatory "*safe harbor*," DHHS has developed a guidance regarding approaches for the 'expert method' to achieve de-identification in accordance with HIPAA. Identifiable healthcare data may be used if they are released by authorization from each individual patient or released as part of a limited data set (LDS). A limited data set has certain identifiers (such as name and street address) removed or masked, but allows dates and more fine grained geo-location than does de-identified data. It may include identifiable information, but only as necessary to complete the proposed research. The recipient of the data must agree to a *data use agreement* (DUA) in which the purpose of the research and proposed uses for the data are described. The DUA also requires securing the data and prohibits re-identification of the information, including linking to other data from the patients. Thus, use of healthcare data from organizations requires both a contractual agreement with the organization, as well as HIPAA compliance with respect to use and disclosure of the data.

Safe harbor: A method of de-identifying health information that involves removing eighteen identifiers from the data before sharing them with an outside party. The identifiers include name, name, address, social security number, phone and fax numbers, email addresses, biometric information and other individually unique information. From 45 CFR 164.514(b)(2); Guidance on the Safe Harbor Method

Data Use Agreement (DUA): A data use agreement is a contractual document for the transfer of PHI that describes the purposes for which the data can be used and prohibits re-identification. From 45 CFR 164.514.

Sample DUA: A template DUA from Harvard Catalyst can be found here [17].

When approaching a healthcare organization for a DUA, a prospective researcher should be prepared to provide a detailed, precise statement of what data elements are required, from what sources, and over what time period. In addition, the investigator must describe how the data will be used and transferred securely to the investigator, and provide a list of all personnel who will be permitted to use the information.

The researcher must also agree that they will:

- Report any unauthorized use or disclosures
- Safeguard the information and describe security measures and how and where data will be stored and protected
- Hold anyone to whom it provides information, e.g., a subcontractor or research collaborator, to the requirements and restrictions of the DUA

Further, the researcher must agree that they will not:

- Further disclose the information, except as permitted by the DUA or as permitted by law
- Contact or re-identify the individuals, or link the data with other datasets that may enable such re-identification

Healthcare facilities often lack resources for providing or transmitting the data to external investigators. Time and resources are required for health IT staff to: 1) collaborate with the research investigator to translate the data requirements into executable queries that will retrieve the data from the enterprise databases; 2) validate the retrieval process or the

data retrieved; 3) provide documentation of how the data were collected, defined, and managed over time at each facility; 4) work with the investigator to come to agreement on file transfer specifications and processes; and 5) answer questions about the data after provision.

Investigators may not be able to estimate the effort required for these activities because they depend on how data are stored and documented; such cost estimation requires conversations with the stakeholders at the healthcare facility. When compensation is required for the provision of data, the DUA is often part of a larger contract for such services. Further, some stakeholders at healthcare facilities may desire intellectual involvement, the right to review research results, or participation in publication of the research relying on data from their institution. Specific details surrounding these types of involvement, restrictions, and rights to review are often included in contracts regarding data provision. Such contracts are substantially more complex when the research requires extensive or ongoing involvement of the stakeholders at the healthcare facility.

DUAs are formal agreements and as such typically require legal involvement from both the providing and receiving organizations. However, many organizations do not provide data for research use on a regular basis and have no established process for DUA consideration or approval. Further, organizational processes for obtaining DUAs and associated contracts are usually not transparent to investigators or others outside the organization. Often, it is difficult to identify an individual who has the authority to consider a draft DUA and present it to the appropriate organizational group(s) for review and ultimate approval. Thus, it is the responsibility of investigators to identify an individual within the healthcare facility who has both the knowledge of the requirements and sufficient authority within the organization to convey the desire for the DUA to the appropriate individuals or groups for consideration. Ideas for initial contacts include an organizational leader with responsibility for research, the Chief Data Officer, or an institutional privacy officer. The informatics department and data warehouse managers are often familiar with the process for developing a DUA and might have examples or template agreements, although the actual agreement will be signed by a designated individual with appropriate authority at the institution. Timelines for working out DUAs between stakeholders at healthcare facilities and external investigators vary greatly; in our experience, intervals range from 6 months to more than 2 years.

Fundamental Differences in Context

There is a fundamental difference in context between data collected retrospectively from EHRs and data collected prospectively for a specific study and according to the study protocol. When data are collected according to a protocol for a research study, the protocol defines the context of the data, for example, "*the second assessment occurs 14 days post baseline.*" The circumstances around data collection, including procedures for taking samples and making observations and recording data, are defined in the protocol, as are other contextual items such as patient positioning, timing, anatomical location. The

resulting data fit exactly into a structured data collection form, i.e., a measurement for each time point taken as prescribed by the protocol. In designing studies, a top-down approach is usually taken starting with the research question and working down to the required data.

By contrast, data captured in an EHR in routine-care settings are from a different context. These data are the result of an individual patient's circumstances and reflect standard procedures at the patient's healthcare facility. In representing the natural course of injury or disease progression and treatment in a patient, such data do not appear in a structured manner. Instead, the structure is one imposed at the facility according to their standards for clinical documentation. These differences suggest that the design process for a clinical study using EHR data should be bidirectional—from the research question down, and from the available data up.

Assessing the Availability of Health Record Data

The content of clinical documentation in the health record is heavily influenced by local documentation practices, the care provided, and methods that are used for billing and reimbursement. Investigators can therefore make few assumptions about what information will be available in an organization's health records. Local workflows, information systems, interests, and procedures can have a substantial impact on data availability and the content and format of clinical documentation. The situations that an investigator will encounter will vary with the organizations from which data are sought. The following scenarios illustrate varying degrees of availability of EHR data.

Scenario 1: Differences in availability of data based on clinics and providers

As part of an initiative to improve the identification and control of chronic hypertension among residents of Durham County, North Carolina, investigators planned to retrieve data directly from EHRs [18]. They anticipated that this project would be feasible because collecting blood pressure in healthcare settings is routine and straightforward, and the relevant health system (Duke University Health System) served an estimated 80 percent of the local population. However, initial queries showed that data on blood pressure were missing from 40% of patient encounters, and further investigation revealed that some specialty clinics (e.g., ophthalmology; orthopedics) did not routinely collect blood pressure. Within the primary care clinics, some providers recorded blood pressure measurements in "free-text" fields rather than the structured data entry fields for systolic and diastolic blood pressure. Other providers took blood pressure measurements but did not document them. Thus, blood pressure data were not as widely available as would have been expected, and the degree of availability varied by clinic and provider.

Scenario 2: Unavailability of historical data due to recent migration to a different information system

We were working with a small group of clinics in a larger study that relied upon access to health record data, and the clinic organization had adopted a new EHR system in the previous year. Historical data from paper charts were not entered into the new system due to cost constraints. All existing patients were scheduled as new patients at their first visit following the adoption of the new EHR, during which they completed the equivalent of a comprehensive intake form. This information was entered for each patient, and the clinical information collected during prior encounters was not available. Even when clinics scanned paper charts as part of the record migration process, the historical information was unavailable to queries on structured data in the EHR.

Scenario 3: Variations in data availability in onsite versus referral providers of services

In a multicenter study relying on EHR data, some internal medicine clinics had onsite endoscopy facilities, and others referred patients to external providers. Among the latter group, some clinics did not incorporate endoscopy results or reports into the EHR, while some entered the result as a coded diagnosis, and others had the endoscopy report scanned in as a document or image file (e.g., PDF). Conversely, clinics with onsite facilities had electronic data available.

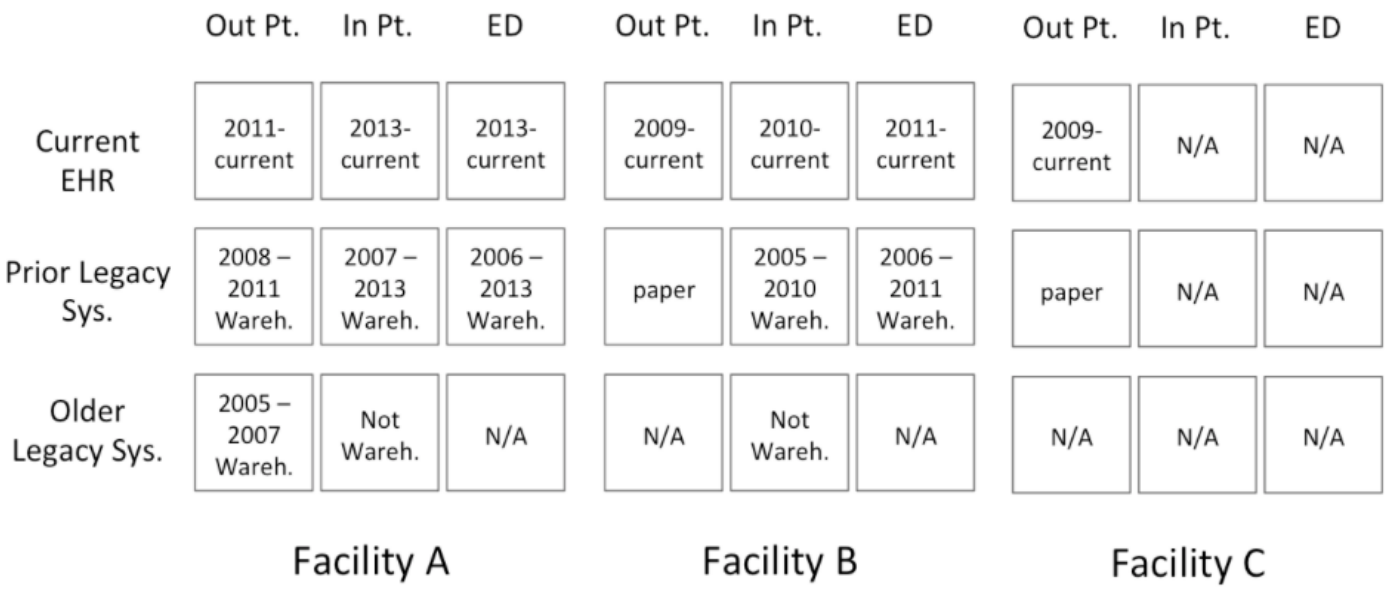
Scenario 4: Variations in outpatient, inpatient, and emergency department data

In the MURDOCK longitudinal community registry, EHR data were collected from regional facilities [19]. The investigators found different degrees of availability for outpatient, inpatient, and emergency department data within the EHRs, as well as variability in the amount and type of data available from *legacy systems* in institutional *clinical data warehouses* (Figure). Investigation of data availability at the broad level of encounter type (outpatient, inpatient, and emergency department) and information system (current, legacy, and older legacy) was informative across all participating sites. For example, the number of "years back" for which data were available across all facilities limited the phenotype algorithms that we could use. The diagrams can be expanded to include differences in data availability between primary care and specialty clinics, or practice management systems in small private clinics not yet using EHRs.

A **legacy system** refers to an older information system for data collection and includes paper-based and electronic systems. These systems were primarily used to store and retrieve data, and were not standardized for interoperability across systems.

A **clinical data warehouse** is "a data base of clinical data obtained from primary sources, such as electronic health records, organized for re-use for secondary purposes" [20].

Figure. Nine-box diagrams of data availability by facility based on encounter type (outpatient, inpatient, and emergency department) and information system (current, legacy, and older health record systems)



Nine-box diagrams display high-level information regarding availability of data by encounter type, information system, and facility.

Describing the availability of data by facility, encounter type, and types of information system (current and historical) in use represents an artificially simple model for most environments, but often provides a useful high-level glimpse of data availability. For example, health systems may have different information systems for each specialty and operational function rather than a common EHR system. Further, systems that support multiple clinics (as in Scenario 1 above) may exhibit variability over time by clinic or even by individual care provider.

Understanding data availability in environments with multiple clinics and providers is considerably more complex than the simplistic approach described above. For example, warehousing of data from legacy systems rarely incorporates all of the data from the older system, and some data elements may be present in the data warehouse while others are not. Understanding data availability for research studies entails up-front, detailed discussions of the facility's collection of each required data element over time, as well as working with the data after it is received to assess variability by clinic, unit, or provider. Further, data obtained from institutional data warehouses have often undergone transformation; these transformations must also be understood by secondary data users.

Understanding the Available Data

There are two key aspects a user of secondary data must understand about the data: the *meaning* and the *quality*. Deficiencies in either can severely affect the data's capacity to support research conclusions or even render data altogether unusable for a study. Therefore, investigators and users of secondary data must understand both meaning and quality of available data prior to using them for research purposes.

Data Meaning

In clinical medicine, a disease, condition, or other concept is often not directly measurable, and its existence is inferred from other measures. For example, diabetes is variously defined by one or more elevated lab measures in a certain period of time. *Operationalization* is the process of defining a concept that is not directly measurable and may create a difference in meaning, for example identifying patients with diabetes through diagnoses, medications, lab values, or some combination thereof. This difference in meaning could make the data inadequate for the needs of the research study, e.g., too broadly or narrowly defined, and can occur either through research operationalization itself or through differences in the healthcare data. There may be a difference between the concept required by the study and either the research operationalization itself, or the meaning of the healthcare data that cause a difference in the operationalization and the desired concept. Some difference may be acceptable as long as it is understood and described. Further, many aspects of how data are defined and processed in the healthcare setting can alter or add undesired variability to the meaning of the data. This can make both the concept and the data used to represent the concept less related, resulting in an increase in *semantic distance.* This is also referred to as representational inadequacy [21], and the potential for increasing the representational inadequacy exists at each step in the handling of data, from origination to the analysis dataset.

Semantic distance represents the distance between the meanings of two messages [22].

Ascertaining what the data mean

Clinical workflows and local clinical documentation procedures can affect not only the availability of data, but also the meaning of the data. After investigators have determined that sufficient data are *available* at a given facility, the next step is to ascertain whether the meaning encompassed by those data will meet the needs of a study. Even within a single site, interpretation of certain treatment and diagnosis codes may differ depending on clinical practices and the individual care provider. For example, data coded as *psychotherapy* versus *psychotherapy with medication management* may return two different populations in settings where clinical psychologists have the ability to prescribe medication versus those where they do not. Most facilities will be able to provide a data dictionary along with data including data element name, data type, valid values for discrete data elements, and field length. However, some facilities may not provide definitions for data elements, the source system/s, the clinical workflow where the data were documented, formulas, or logic for calculated data elements.

A definition that conveys the meaning of the data element will help a researcher understand if the values have the same meaning across the institution and over time, and, in a multicenter study, how comparable they are to data from other sites. Unfortunately, simple data element descriptions included in information system data dictionaries usually do not provide this type of "meaning" or information. In a recent project that relied on EHR data from multiple facilities, the investigators worked with each facility through in-person and telephone interviews to obtain as much of the information depicted in Table 1 as possible, with the goal of ensuring correct interpretation and supporting later analysis of the data.

Table 1. Information gathered to determine the meaning of a data element and assure correct interpretation in later analysis

Information	Examples
Source system/s for each data element	Bed side monitor, local lab, or central lab
How close the source system is to the original point of charting	Lab value from bedside monitor charted at the bedside
Source of the data values	Clinician, patient, both, or instrument
Uniformity of the clinical work flow in which a data element was collected	Institutional policy for lab value from bedside monitor to be charted at the bedside
How the capture of the data has changed over time	Bedside monitors were first used on the unit in 2005, all prior data are from the facilityu2019s on-site lab
How relevant ICD and CPT codes are applied within the facility	Multiple Universal Laboratory Order Codes (LOINC) codes exist for the same laboratory data element when the same measurement is performed on different samples, e.g., plasma versus whole blood, or results are

Table 1. Information gathered to determine the meaning of a data element and assure correct interpretation in later analysis

Information	Examples
	reported in different units, or results are obtained using different analytic methods. These are differentiated by different LOINC codes in the data.
Consistency of charting the data element across the facility	Whether some values for a data element are sometimes skipped or documented in text fields
Whether the data element contains both data from devices and manual measurement and if so how these are differentiated	Bedside monitors were first used on the unit in 2005, all prior data are from the facility (2019) on site lab
Any cleaning, standardization or other transformations performed on the data element	Values flagged as originating from hemolyzed samples are excluded from the warehouse

As of this writing, few healthcare facilities will have the information in Table 1 on hand. Obtaining this information will likely require intensive interview sessions with personnel from the healthcare facility who understand the data systems and the clinical workflow used to record patient data in medical records.

Data Quality

Another way that data can fall short of meeting study needs is through information loss and degradation. Information can be lost through data reduction, disassociation from context, or error. Information is degraded through introduction of errors as data during collection or processing (for instance, a keystroke or transcription error, or selecting the incorrect item from a menu). Secondary-use data can be subject to many processing steps both at the healthcare facility and after receipt by the investigator (i.e., transcription, coding, and data transformation). Each time data values are accessed in circumstances that allow changes to be made to the data, the possibility of introducing error, information loss, and/or degradation is present. These can change not only individual data values, but also the distributional properties of a dataset, and they can generate spurious outliers or increase the seeming variability of the dataset.

Assessing data quality

Assessing the quality of the data is a key part of gauging their suitability for a study. Data quality is a multifaceted concept, and characteristics such as accuracy, completeness, timeliness, traceability, etc., may be important to a given study. But although many possible characteristics can be named, we would advise researchers to focus on those characteristics that are most relevant to supporting research conclusions for a given study.

The two characteristics that we have found to be most important in our previous work are consistency and completeness.

Consistency

Consistency can be thought of in two parts: external and internal. External consistency involves a comparison to an external source of information that is independent and has some expectation of accuracy, for example, comparison to gold standard or comparison to relevant data from a different or upstream source. The closer the gold standard is to truth, the closer the external consistency assessment is to an actual assessment of accuracy. Internal consistency is a comparison to other available data within the same dataset, i.e., from the same patients, across clinical or research sites, or over time. Such internal consistency checks include programmed logic checks for procedure dates occurring after a death date, comparisons of weight and height over time, or comparison of coded values between comparable sites in a multicenter study.

Completeness

Conceptually, completeness is the presence of necessary data. The four mutually exclusive elements of a comprehensive assessment of completeness presented below were adapted from recent theoretical work by Weiskopf et al., [23], and can also be found in the tool *Assessing Data Quality for Health Systems Data Used in Clinical Research* [9].

1. **Data element completeness** refers to whether or not all the necessary variables in a candidate dataset are present. For example, "Are the right 'columns' present?" Data element completeness is assessed by examining metadata, such as a data dictionary or list of data elements contained in a dataset and their accompanying definitions, and comparing this information against the variables required in the analytic or statistical plan. With adequate data documentation, data element completeness can be assessed without examining any data values.

2. **"Column" data value completeness** refers to the percentage of data values present for each data element. Note, however, that often (as in normalized structures) more than one data element may be stored in a database column. When data are *normalized* in this context, it means that multiple descriptions are allowed within a data column if they have the same meaning, for example, acute myocardial infarction means the same thing as heart attack or cardiac arrest [24]. The word *column* is used to help the reader visualize the concept because normalized data structures are often flattened to a 1-column-per-data-element format to generate and report data quality–related statistics. Column data value completeness is assessed by structuring the dataset in a "1-column-per-data-element" format and calculating the percentage of non-missing data for each column, with non-missing defined as "not null and not otherwise coded to a null flavor." Null flavors (e.g., not applicable, not done) are defined in the International Organization for Standardization (ISO) 21090 [25] and Health Level Seven International (HL7) [26] data type definition standards.

3. **Ascertainment completeness** refers to the percentage of eligible cases present; i.e., "Do you have the right 'rows' in the dataset?" Ascertainment completeness usually cannot be verified with absolute certainty because assessment options are typically based on a comparison to a subset of the data or a similar population. These include but are not limited to: 1) chart review in a representative sample and 2) comparison to one or more independent data sources covering the same population or a subset of that population. Ascertainment completeness is affected by data quality problems, by phenotype definition and execution, and by factors that bias membership of a dataset. Other issues commonly evaluated in an ascertainment assessment include the presence and extent of duplicate records and records for patients that do not exist (e.g., an error in the medical record number creates a new case; a patient gives a name other than his or her own), or duplicate events such as a single procedure being documented more than once. Ascertainment completeness and phenotype validation significantly overlap in goals and can be accomplished together.

4. **"Row" data value completeness** refers to the percentage of cases/patients with sufficient data values present for a given data use. The presence of data values in rows is assessed using study-specific algorithms programmed to calculate the percentage of cases with all data or with study-relevant combinations of missing and non-missing data (e.g., in the case of body mass index [BMI], the percent missing of "either weight OR height" might be calculated, because missing either data point renders the case unusable for calculating BMI).

A comprehensive completeness assessment consists of all four of the components described above. In terms of effort, column completeness is accomplished through a review of data elements available in a data source, as described in the understanding the availability of data section. Evaluating column data value completeness and row data value completeness are straightforward computational activities. Evaluating ascertainment completeness, however, can be a resource-intensive task because it may involve activities such as chart review on a representative sample, or electronic comparisons among several data sources.

Identifying Populations and Outcomes of Interest

Healthcare facilities are obligated by federal regulations to request only the "minimum necessary" data elements to answer planned research questions. Thus, investigators must specify the patients from whom data are required as well as the needed data elements; this specification is commonly referred to as a *phenotype*.

Phenotype: "a clinical condition or characteristic that can be ascertained via a computerized query to an EHR system or clinical data repository using a defined set of data elements and logical expressions [27]."

Specifying the needed data is not an easy task for investigators. There are usually multiple possible definitions for a condition of interest, and an investigator must determine a definition that can be applied with the required sensitivity and specificity given the available data. This requires identifying or developing definitions and evaluating them for use on the intended study.

In addition, investigators may not be familiar with the content, format, and structure of the EHR data at a given healthcare facility. To look at it another way, specifying the data of interest is like describing a book that you've never read. Investigators without knowledge of the EHR data content, format, and structure are forced to conceptually describe the needed data, leaving data analysts at the healthcare facility to apply the conceptual definition to the available data.

Consider a data request for a hypertension control study: the conceptual definition may include, "patients for whom three blood pressure measurements within a 1-year period are greater than 140/90 mm Hg or who have a diagnosis of hypertension, or who are prescribed a blood pressure lowering agent." An analyst writing computer code to access these data will need additional information, including:

- Which of 51 ICD-9-CM codes for hypertension should be included?
- Should outpatient, inpatient, and emergency department encounters be included?
- Should automated blood pressure monitoring data be included?
- Should orders, medication reconciliation, and fulfillment data be used?
- How far back in time should data be evaluated?
- Should rolling year versus calendar year be used?
- Should deceased patients be included?
- Should perioperative data be included?
- Should hypertension in the gestational period be included?

The analyst will need to further apply codes used in the facility data as well as system variables such as dates and time stamps. Accordingly, after a phenotype definition is chosen, some further discussion is typically required to fit the definition to the health system data.

Guidance is available from the NIH Health Care Systems Research Collaboratory on finding, evaluating, developing, and testing phenotype definitions [27].

Record Linkage Considerations

Some studies use data from multiple records and sources. This requires matching (or combining) data while ensuring that they refer to the correct patient. Such matching is commonly called *record linkage*, or more formally, entity resolution. To perform record linkage, a study receiving data from multiple healthcare facilities matches data from each facility to the correct patient. Likewise, patient-reported outcomes (PROs) or home

monitoring data may also need to be linked to the patients' EHR. Some studies may use fully identified data for record linkage while others may use de-identified data, which may require linkage with other de-identified data, such as Medicaid or Medicare data. Depending on the data sources required, some studies may use both identified and de-identified data.

De-Identified Data (From 45 CFR §164.514) De-identified data refers to "Health information that does not identify an individual and … there is no reasonable basis to believe that the information can be used to identify an individual." Identifiers include: name, geographic location and zip codes, dates (birth, admission, discharge, etc.), telephone numbers, social security numbers, email addresses, medical record numbers, health plan beneficiary numbers, account numbers, license numbers, vehicle identifiers, full face photographic images, etc.

Probabilistic and deterministic methods for fully identified and de-identified record linkage are well described elsewhere [28]. Fully identified record linkage uses data elements with a high degree of completeness and uniqueness such as name, address, telephone number, email address, and Social Security number. De-identified record linkage uses data elements with high completeness that in combination provide increased uniqueness over that of any single data element; e.g., procedure date, encounter date, procedure code, encounter facility, sex, *and* age. Aside from the differences in data elements used for matching, the methodology for matching and linking is similar.

When records need to be linked for secondary use of EHR data, there may be extra steps and special considerations. For example, a study may require quantifying false positives and false negatives in the process of record linkage, and doing so requires creation of a truth set and testing against it. When direct measurement is not possible, relative methods are used. For example, in order to provide only the minimum necessary information, large healthcare facilities may perform record linkage in-house, and an outside researcher may have no control over the record-linkage method used.

In addition, facilities are not usually able to provide evaluative measures of false positives and false negatives. To test for data quality in this situation, we provide record linkage software configured with an evaluated algorithm (a "black box") and ask the personnel at healthcare facilities compare the black-box results with their record linkage data. This methodology demonstrates how different their results are from our characterized algorithm. In addition, when the healthcare facility links the data, the matching is reduced to a simple deterministic match by the study team after the data are received.

Simple deterministic match: In a simple deterministic match, two records are said to match if all or a combination of specified identifiers are identical. For identified data, social security numbers can serve as unique identifiers. A combination of other de-identified data elements could also be used.

For studies linking prospectively collected data to EHR data such as PROs or home monitoring data, fields that will be linked are considered and planned for in the data collection phase so that the matching by the study team is reduced to a simple deterministic match. Any prospective data collection should provide for record linkage needs as part of the data collection plan.

Record linkage results can be very sensitive to data quality in terms of those slight differences in words and phrases known as *lexical variation*, as well as case differences, spaces, odd and control characters, unexpected patterns (such as 5-digit versus long zip codes, or dashes, dots, and parentheses in phone numbers. Thus, data standardization should be considered along with record linkage and discussed with personnel in facilities linking their own data.

If a limited dataset is provided for record linkage, a DUA that specifies all intended linkages will be needed. This document will be reviewed by the providing facility's IRB to ensure adherence to the requirements for limited datasets, including the provision that no attempt be made to re-identify data.

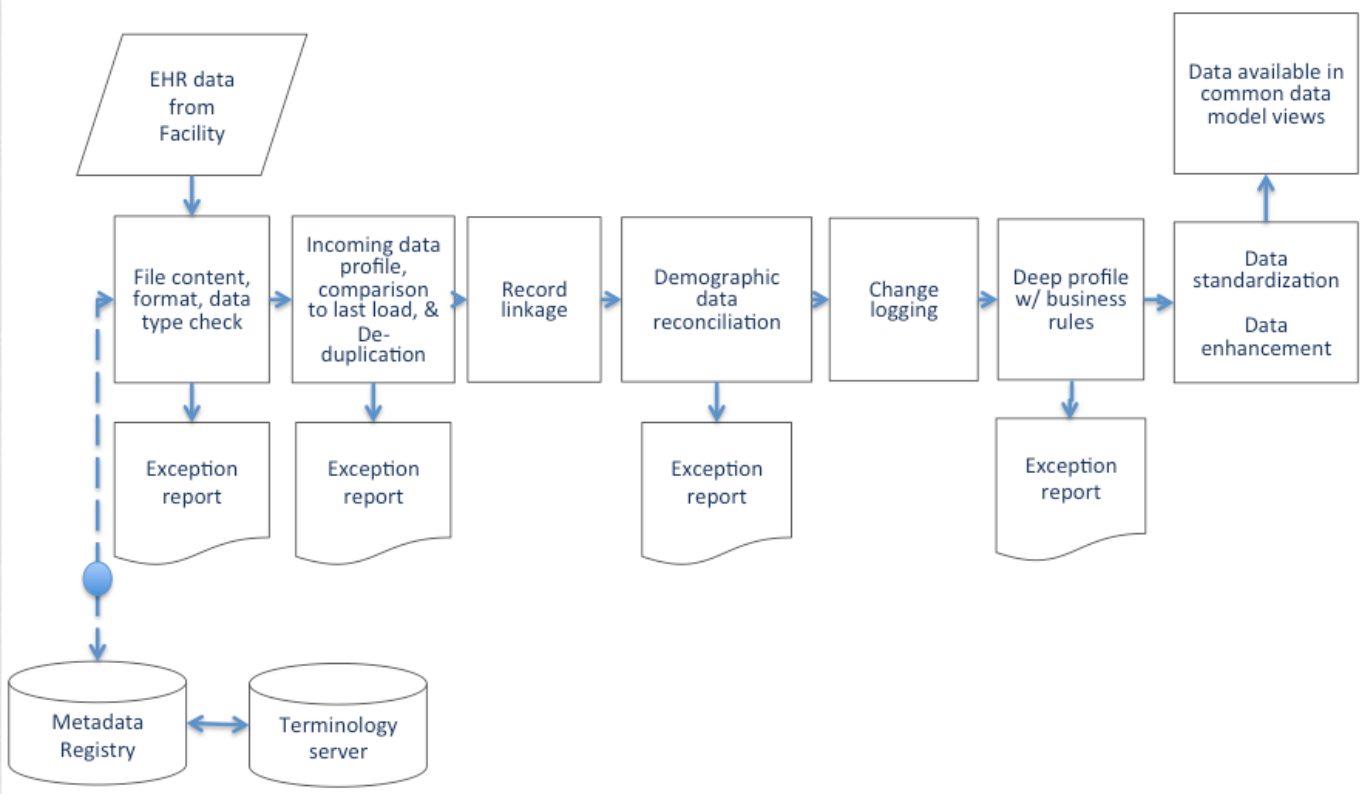
Managing EHR Data Obtained for Research

A *federated database* is a virtual database that functions as a composite of all the databases in the system. Use of federated data—in which a query is sent to the database rather than the data being sent to the investigator—is in many cases the preferable option where such capabilities exist. If this is not possible, the investigator must undertake a significant amount of work to receive, process, and manage health record data. Either way, the investigator is responsible for understanding and demonstrating the ability of the data to support research conclusions (through data quality measurement and impact assessment) and for assuring reproducibility. Reproducibility is accomplished by demonstrating *traceability* from data origination to the analysis dataset. When federated data are used, traceability information should be obtained from the data holder. If the investigator alters data, e.g., cleaning, standardizing or other transformations, the investigator must provide traceability documentation. Such traceability information should enable reproduction of the analysis dataset from the original raw data. The remainder of this section will cover data transformations commonly performed on health record data by secondary data users. Below, we will describe common methods for maintaining traceability.

Traceability is the ability to verify the origin of and all changes to the data.

The Figure shows common data transformations undertaken in the receipt and processing of health record data. Some of the steps in the Figure show discrepancy detection and imply seeking resolution. Such data-cleaning techniques are usually not necessary with secondary data use, but they may be needed in some studies, and we have included them for completeness. Throughout data processing, maintaining traceability means documenting any change to each data value. Opportunities for data changes to secondary data fall into two categories: changes made by the data provider, and changes made by the data recipient. Maintaining traceability requires documenting both. Failure to document changes to data can result in inability to answer questions about the analysis, or worse, can make research findings essentially not reproducible.

Figure: Transformations commonly performed on health record data by secondary data users



Data are received from providing healthcare facilities. There is typically a file content, format, and data type check to assure that the files conform to agreed specifications and can be further processed. Any problems are reported to the providing facility. A second step may include data profiling, comparisons to the last file received, and de-duplication to catch and report any problems. After files and the contained data are deemed as expected,

incoming records are linked with existing data or data received from other sources. Following linking, other data elements not used in linkage are reconciled to confirm the matches. Any problems are reported and resolutions sought. Following linkage reconciliation, a change-logging step may occur where new data received are compared to existing data (previously received from the facility) and any data changes are logged. Changes to data made at healthcare facilities may be of interest to the study, especially if reported results change over time due to facility-based data changes. Often deeper data consistency and completeness checks are performed with exceptions reported and discussed with the providing facility. After data have been loaded, linked, and checked, data standardization and enhancement can be undertaken. Finally, if a *common data model* is used, data are made available through the common data model.

Common data model is a way of organizing data into a standard structure.

Changes Made by the Data Provider

File content and format checks do not usually result in the secondary user changing data; if problems in these areas are detected, they are reported to the data provider and a new file is sent. It is advisable to maintain a record of all files received and all problems reported because this history provides documentation of the condition of the data as received. The secondary data user is responsible for all changes to data after receipt; thus, documenting the original state of data is important.

When secondary users check data, e.g., through data profiling, business rules, or other consistency and completeness checks, data discrepancies may be detected. Because the secondary data user is not the origin of the data, these discrepancies can be used as an assessment of the data, or in some cases reported back to the provider of the data for resolution, who is typically best placed to make any needed changes. In such a circumstance, the provider may send a new data transfer with all of the data or just the corrections. Records of the reported discrepancies should be kept. Corresponding data changes made by the data provider will be captured in the change-logging step (Figure 2). Not all discrepancies will prompt a data change; thus, to prevent replicative reporting of a discrepancy, the receiving data system should log that a discrepancy has been reported so it will not be reported to the data provider again.

Changes Made by the Data Recipient

Data standardization is the only process in which the secondary user actually makes changes to the data. Data standardization can vastly improve the usability of the data through measures such as:

- Coding free-text data with standard controlled terminologies, including ICD coded diagnoses, Universal Laboratory Order Codes (LOINC) coded labs, and RxNORM coded medications
- Converting local or institutional codes to standard controlled terminology
- Adding identifiers such as the National Provider Identifier, which is a unique identifier for all healthcare providers as part of the National Plan and Provider Enumeration System (NPPES)
- Imputing missing or invalid data
- Cleaning up text fields such as names, addresses, and telephone numbers so that the data exhibit consistent pattern and case.

The best practice in the case of data standardization is for all changes to be recorded; many secondary users accomplish this by retaining the original data values as received and creating a second data element to hold the transformed values. In this case, the date of the transformation is also documented, and both of these become value-level metadata.

Lastly, while most will likely be computed, some transformations, such as medical record abstraction, coding, or clinical interpretation of diagnostic testing, may be performed manually. Specifications for manual and computational transformations should be created and maintained. Such specifications will help describe data handling and may be evaluated by subsequent researchers for appropriateness. *Inter-rater reliability* of manual steps should be measured and confirmed. Programmatic algorithms for transformations should be tested to confirm that they are working as specified. In reproducibility, the fidelity of the transformation is as important as documentation that it took place.

Inter-rater reliability: the degree of agreement among *raters* or those doing the manual transformations.

Archiving and Sharing Data After a Study

Data are archived for several purposes, including to enable auditing for quality assurance and regulatory compliance, to ensure the reproducibility of studies performed with the data, or to answer other questions about the research. Best practices for data archival are very similar to those that support data sharing, and different regulations specify different intervals for retaining records. The National Institutes of Health requirements for data sharing plans (2003) are relatively new. They require that federally funded studies receiving over $500,000 per year have a data sharing plan describing how data will be shared, that shared data be available in a usable form for some extended period of time, and that the least restrictive method for sharing of research data should be used, provided it maintains scientific integrity and appropriate protection for participants and health systems.

Key points in the NIH policy and guidance include:

- The privacy of participants should be safeguarded.
- Data should be made as widely and freely available as possible.
- Data should be shared no later than the acceptance for publication of the main study findings.
- Initial investigators may benefit from first and continuing use of data, but not from prolonged exclusive use.

The Health Care Systems Research Collaboratory Data Sharing Policy and Policy Considerations (2014) document describes additional considerations for sharing data originating from routine clinical care and used in pragmatic clinical research. Because individual participant consent may be waived by the IRB in accordance with the federal regulations (45 CFR part 46) in some pragmatic trials, special considerations apply. For example, researchers are only allowed to use and store data elements specifically authorized for research use, either by participant consent or by formal waiver of consent by the IRB [29]. Further, investigators are only expected to share the specific data elements on which their analyses are based. The detailed, original data would not need to be retained by investigators to fulfill data sharing requirements [29]. Precautions to protect healthcare systems and providers who collaborate with investigators for pragmatic research are also necessary, such as allowing data sharing through a restricted data enclave that limits access to researchers [29].

Effective data sharing requires more than just making the data available to others. Information about the data (metadata) is required for others to understand and appropriately use shared data. One example of this is provided by the National Institute for Drug Abuse (NIDA) Clinical Trials Network (CTN), which hosts a web-based data-sharing hub that facilitates sharing research findings from all clinical trials conducted by the network (https://datashare.nida.nih.gov). Although the NIDA CTN data share does not contain any clinical studies using EHR data, it shows that it's possible to make data and supporting documentation publicly available while at the same time protecting privacy and confidentiality of research participants. The following documentation is provided with the dataset from each trial: 1) the study protocol, 2) reference to study publication of primary outcome, 3) data sets (SAS and ASCII), 4) annotated data collection forms, 5) a data dictionary defining each data element, and 6) study-specific de-identification notes. Together, this information provides crucial context, such as where the data originated and how they were collected and analyzed. Each dataset is provided in the data model used for the trial as well as the Clinical Data Interchange Standards Consortium (CDISC) standard Study Data Tabulation Model (SDTM) in both SAS transport file and ASCII format. Further, the information about the NIDA CTN trials on ClinicalTrials.gov links to the data share site.

Data sharing does not have to be as public or detailed as in the approaches described above. Datasets and supporting documentation may be shared as on-line supplemental material in journals, university websites, or sites maintained by research groups or investigators. Some data sharing plans merely employ a reference contact in the primary publication.

Bibliography

1. Kush R, Alschuler L, Ruggeri R, et al. Implementing Single Source: the STARBRITE proof-of-concept study. J Am Med Inform Assoc 2007;14:662–673. PMID: 17600107 PMCID: PMC1975790. doi: 10.1197/jamia.M2157.

2. Koran LM. The reliability of clinical methods, data and judgments (first of two parts). N Engl J Med 1975;293:642–646. PMID: 1097917. doi: 10.1056/NEJM197509252931307.

3. Koran LM. The reliability of clinical methods, data and judgments (second of two parts). N Engl J Med 1975;293:695–701. PMID: 1160937. doi: 10.1056/NEJM197510022931405.

4. Van der Lei J. Use and abuse of computer-stored medical records. Methods Inf Med 1991;30:79–80. PMID: 1857252.

5. Burnum JF. The misinformation era: the fall of the medical record. Ann Intern Med 1989;110:482–484. PMID: 2919852.

6. Meads S, Cooney JP. The medical record as a data source: use and abuse. Top Health Rec Manage 1982;2:23–32. PMID: 10255780.

7. RFA-RM-13-012. NIH Health Care Systems Research Collaboratory - Demonstration Projects for Pragmatic Clinical Trials Focusing on Multiple Chronic Conditions (UH2/UH3). 2013. Available at: http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-012.html. Accessed February 10, 2015.

8. Kahn M, Brown M, Chun A. DQC White Paper: A consensus-based data quality reporting framework for observational healthcare data. Submitted to eGEMS Journal, December 2013.

9. Zozus MN, Hammond WE, Green BB, et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. Available at: http://sites.duke.edu/rethinkingclinicaltrials/assessing-data-quality/. Accessed February 10, 2015.

10. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010;67:503–527. PMID: 20150441. doi: 10.1177/1077558709359007.

11. Braun BI, Kritchevsky SB, Kusek L, et al. Comparing bloodstream infection rates: the effect of indicator specifications in the evaluation of processes and indicators in infection control (EPIC) study. Infect Control Hosp Epidemiol 2006;27:14–22. PMID: 16418981. doi: 10.1086/498966.

12. Williams SC, Watt A, Schmaltz SP, et al. Assessing the reliability of standardized performance indicators. Int J Qual Health Care 2006;18:246–255. PMID: 16431865. doi: 10.1093/intqhc/mzi098.

13. Watt A, Williams S, Lee K, et al. Keen eye on core measures. Joint Commission data quality study offers insights into data collection, abstracting processes. J AHIMA 2003;74:20–25; quiz 27–28. PMID: 14618843.

14. Gibbs D. For Debate: 250th anniversary of source document verification. 1996;313:798–798. doi: 10.1136/bmj.313.7060.798.

15. Zozus MN, Pieper C, Johnson C, et al. Factors impacting accuracy of data abstracted from medical records.2015;10:e0138649. doi: 10.1371/journal.pone.0138649. PMID:26484762.

16. International Committee of Medical Journal Editors. Protection of Research Participants. 2015. Available at: http://www.icmje.org/recommendations/browse/roles-and-responsibilities/protection-of-research-participants.html. Accessed October 28, 2015.

17. Harvard Catalyst Data Protection Subcommittee. Harvard Catalyst Data Use Agreement for Limited Data Sets. 2014. Available at: https://catalyst.harvard.edu/pdf/regulatory/Harvard_Catalyst_Template_LDS_DUA.pdfAccesed July 2, 2020.

18. Thomas KL, Shah BR, Elliot-Bynum S, et al. Check it, change it: a community-based, multifaceted intervention to improve blood pressure control. Circ Cardiovasc Qual Outcomes 2014;7:828–834. PMID: 25351480. doi: 10.1161/CIRCOUTCOMES.114.001039.

19. Bhattacharya S, Dunham AA, Cornish MA, et al. The Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study Community Registry and Biorepository. Am J Transl Res 2012;4:458–470. PMID: 23145214.

20. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer; 2013.

21. Tcheng J, Nahm ML, Fendt K. Data quality issues and the electronic health record. 2010;2:36–40.

22. Cooper MC. Semantic Distance Measures. 2000;16:79–94. doi: 10.1111/0824-7935.00106.

23. Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46:830–836. PMID: 23820016 PMCID: PMC3810243. doi: 10.1016/j.jbi.2013.06.010.

24. Wolters Kluwer Health. Data Normalization: The Foundation of Forward-Thinking Initiatives. 2014. Available at: http://cdn2.hubspot.net/hub/208942/file-2026393142-pdf/Data-Normalization-Final_11.4.pdf?t=1415113836361&t=1423751735590.

25. International Organization for Standards. ISO 21090. Health Informatics – Harmonized Data Types for Information Interchange. 2011. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=35646.

26. Health Level Seven International. HL7 Data Type Definition Standards. Available at: http://www.hl7.org/implement/standards/product_section.cfm?section=2&ref=nav. Accessed Nov 4, 2015.

27. Richesson RL, Smerek M. Electronic Health Records-based Phenotyping, in Rethinking Clinical Trials A Living Textbook in Pragmatic Clinical Trials. NIH Health Care Systems Research Collaboratory. Published June 27, 2014. Available at: http://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/. Accessed February 10, 2015.

28. Talburt JR. Entity Resolution and Information Quality. San Francisco, Calif: Morgan Kaufmann/Elsevier; 2011.

29. NIH Health Care Systems Research Collaboratory Data Sharing Policy Considerations document V.1, updated June 23, 2014. Available at: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Collaboratory_DataSharingPolicy_Considerations-Document_June232014.pdf. Accessed February 10, 2015.