Data Quality and Integration Issues in Electronic Health Records

Chapter · December 2009

DOI: 10.1201/9781420090413-c4

CITATIONS
15

READS
557

6 authors, including:

Ricardo João Cruz-Correia
Center for Health Technology and Services Research (CINTESIS)
174 PUBLICATIONS **1,363** CITATIONS
SEE PROFILE

Pedro Pereira Rodrigues
University of Porto
133 PUBLICATIONS **1,675** CITATIONS
SEE PROFILE

Alberto Freitas
Center for Health Technology and Services Research (CINTESIS)
124 PUBLICATIONS **347** CITATIONS
SEE PROFILE

Rong Chen
Karolinska Institutet
26 PUBLICATIONS **326** CITATIONS
SEE PROFILE

4

Data Quality and Integration Issues in Electronic Health Records

Ricardo João Cruz-Correia, Pedro Pereira Rodrigues, Alberto Freitas, Filipa Canario Almeida, Rong Chen, and Altamiro Costa-Pereira

CONTENTS

4.1 Introduction

Thirty years ago, Komaroff [1] warned that medical data collected on paper records were defined and collected with a marked degree of variability and inaccuracy. He claimed that the taking of a medical history, the performance

of the physical examination, the interpretation of laboratory tests, even the definition of diseases, was surprisingly inexact [1]. A decade ago, Hogan and Wagner [2] argued that electronic health records (EHRs) were not properly evaluated regarding data accuracy. Some studies have been published meanwhile showing that many of the problems found by Komaroff can still be found today.

EHRs are usually used for purposes other than healthcare delivery, namely, research or management. This fact has an important impact on the manner in which data are introduced by healthcare professionals, on how data are recorded on databases, and also on the heterogeneity found when trying to integrate data from different Information Systems (IS). This chapter focuses on DQ and data integration issues when using EHRs on research. The sections of this chapter describe the potential problems of existing EHRs, how to detect them, and some suggestions to overcome them. Some original studies are included in this chapter and are identified as case studies.

4.2 Fundamental Concepts

To better understand the issues covered in this chapter, it is important to understand how patient information flows in healthcare, what are IS (such as the EHR), what is the nature of health-related information, and who are the different actors involved in the process.

4.2.1 The Information Flow

Healthcare is information- and knowledge -driven. Good healthcare depends on taking decisions at the right time and place, according to the right patient data and applicable knowledge [3]. Communication is of utmost relevance in today's healthcare settings, as health-related activities, such as delivery of care, research, and management, depend on information sharing and teamwork [4].

Providing high-quality healthcare services is an information-dependent process. Indeed, the practice of medicine has been described as being dominated by how well information is processed or reprocessed, retrieved, and communicated [5]. An estimated 35% to 39% of total hospital operating costs has been associated with patient and professional communication activities [6]. Physicians spend over a quarter [7, 8] and nurses half [9] of their time writing up patient charts.

A patient record is a set of documents containing clinical and administrative information regarding one particular patient, supporting communication and decision-making in daily practice and having different users and purposes [10]. It exists to memorize and communicate the data existing on a particular individual, in order to help in the delivery of care for this person.

Records are not only an IS but also a communication system that enables communication between different health professionals and between the past and the present [11, 12]. Patient records, the patient, and published evidence are the three sources needed for the practice of evidence-based medicine [3]. They are used for immediate clinical decisions (either by the author or by others), future clinical decisions, quality improvement, education, clinical research, management, and reimbursement, and to act as evidence in a court case.

In practice, quite frequently patients are incorrectly registered or data items can be inaccurately recorded or not recorded at all. The quality of patient data in computer-based patient records has been found to be rather low in several health IS [2, 13, 14]. Most sources of poor DQ can be traced back to human error [13, 15] or bad system design [14]. Moreover, the assessment of the correctness of collected patient data is a difficult process even when we are familiar with the system under which it was collected [16].

Figure 4.1 represents the information flow diagram from patient observation to the use of data by the researcher (based on Hogan and Wagner [17] and Savage [18]). In each step of the information flow (represented by the arrows in the figure), it is possible to encounter problems resulting in data loss or data misinterpretation. It becomes obvious that there is a long way to go from the actual patient data to the data used by the researcher. It should be emphasized that, currently, we just do not have people who are entering information into a computer, we also have computers entering data into each other [19].

4.2.2 What are Models and IS?

To better understand the relation between human artifacts (e.g., IS) and the real world (e.g., healthcare delivery) a description of the steps needed to create a human artifact is presented. These steps stress the fact that the quality of the artifacts is highly dependable on the quality of several models:

1. Create a model of the real world with whom the artifact will interact (e.g., information flow in the obstetrics department).
2. Create a model of the artifact (e.g., Unified Modeling Language diagrams describing the behavior of the obstetrics department IS to be implemented).
3. Implement the artifact based on the model (e.g., implement an Obstetrics Patient Record to be part of the information flow in the obstetrics department).
4. Use the implemented artifact on the real world (e.g., use the implemented Obstetrics Patient Record on a daily basis on a particular obstetrics department).

Mathematical and computational models exist to describe behaviors, that is, to explain how a real-world system or event works. They simplify or ignore

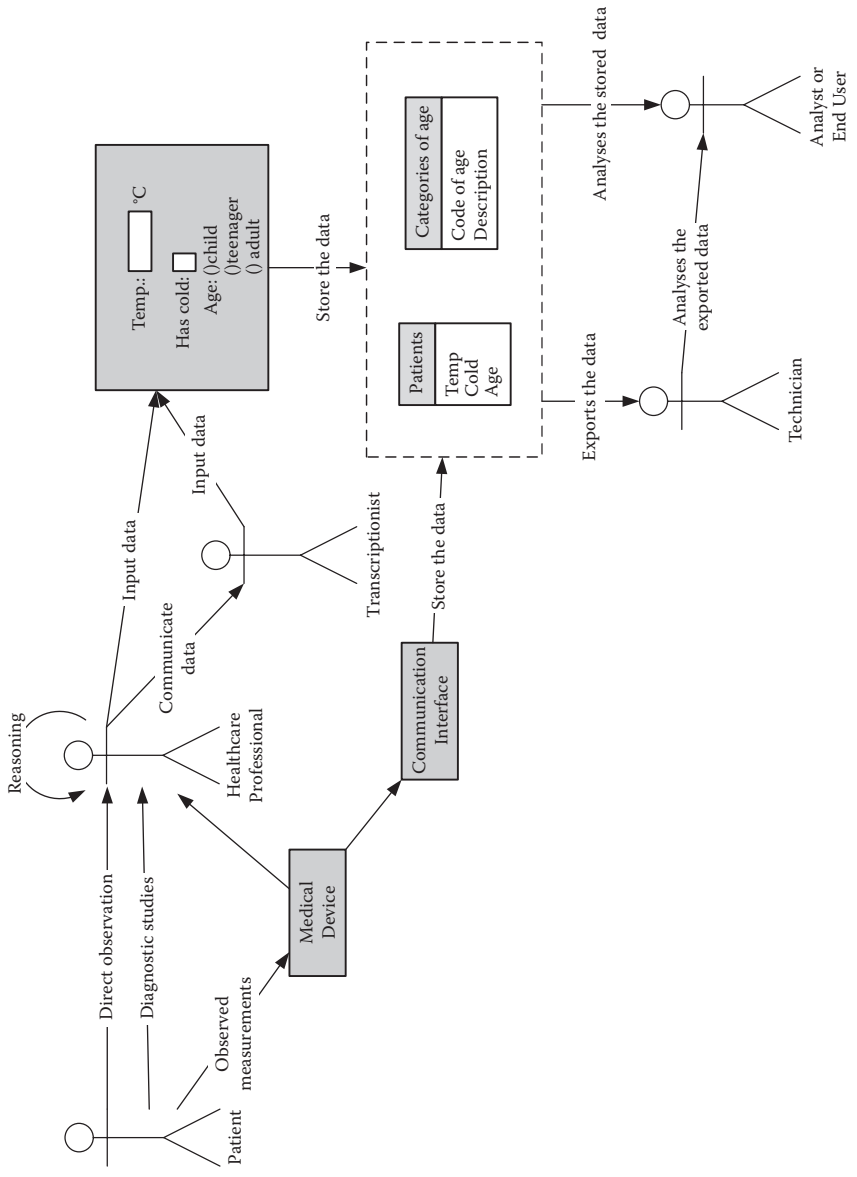

FIGURE 4.1
Information flow diagram from patient observation to the use of data by the researcher.

some details in order to make them more understandable. Generally, they allow complex systems to be understood and make their behaviors predictable. Models can be presented as equations, diagrams, or schemes. They can give false descriptions and predictions when used in situations they were not planned for.

IS are models and being so, they are wrong, that is, they ignore some details not needed for the purpose of the IS. The problem arises when the data collected by the IS are used for different purposes than the ones it was designed for, and for new purposes that ignored details have relevance.

A data model is a plan for building a database. Data modeling produces a formal description of the data that represent concepts of interest to a specific domain (people, places, etc.), and indicates how entities are conceptually related to one another. Data modeling in healthcare is a difficult and time-consuming task due to the vastness of the domain, the complexity of the knowledge, and the wide variety of participants, all with slightly differing views about the process [20]. In this context of healthcare systems, databases must address two important requirements: rapid retrieval of data for individual patients and adaptability to the changing information needs of an institution [21]. There are four basic steps to generic data modeling for a clinical IS: to develop a detailed schema of the medical data, to filter out concepts and relations that do not vary across patient records, to transform the detailed schema into a generic schema, and to implement the generic schema using a database management system [20].

An introduction to healthcare IS can be found in Chapter 2. It focuses on Extensible Markup Language (XML)–based representation of EHR, which is the most promising future representation. However, most existing EHR systems still use the relational model, which stores data in tables, with rows and columns (columns are also called attributes). Even though some parts of our discussion assume a relational model for EHR, the same principles also apply for XML-based EHR representation.

Another important lesson is that it is difficult to propose which data must be structured and which terms must be offered for data entry, before knowing the purpose for which data will be used. In prospective studies, the study is designed before the data are collected, whereas retrospective studies rely on data already collected for a different purpose. Table 4.1 presents some of the differences between retrospective and prospective studies: (1) the time when the meaning of each data fields is known to the researcher, (2) the existence of a research protocol for data collection, (3) the control on the types of users, and (4) the purpose for data collection. Clinical research will often require data that have a high granularity and are recorded uniformly, which will not always correspond to the format in which data are recorded for patient care. Completeness, accuracy, and required uniformity of data, therefore, remain functions of the use of the data. Nevertheless, we must conclude that structuring narrative data does not per se guarantee a thoroughness

TABLE 4.1

Retrospective versus Prospective Studies

Retrospective	Prospective
1. Researcher has to find meaning of each data field	Researcher sets the meaning of each data field
2. Data are collected without protocol	Protocol is predefined
3. Users are heterogeneous	Users are controlled
4. Data are collected for different purposes	Data are collected for research

and retrievability of routinely recorded clinical data for subsequent use in clinical research [22].

4.2.3 The Nature of Health Information

Information is described as the interpretation of data and knowledge that intelligent systems (human and artificial) perform to support their decisions. Data are central to all medical care because they are crucial to the process of decision-making. Data provide the basis for categorizing the problems a patient may have or identifying a subgroup within a population of patients. All medical care activities involve gathering, analyzing, or using of data. They also help the physician to decide what additional information is needed and what actions should be taken to gain a greater understanding of a patient's problem or to treat more effectively the disease that has been diagnosed [23]. Health informatics can help doctors with their decisions and actions, and improves patient outcomes by making better use of information, making more efficient the manner in which patient data and medical knowledge is captured, processed, communicated, and applied.

Health-related data are not homogeneous in nature. They range from narrative, textual data to numerical measurements, recorded biological signs, and images.

Narrative data accounts for a large component of the information that is gathered in the care of patients. They include the patient's symptoms, his/her description of the present illness, medical history, social and family history, the general review of systems, and physical examination findings. Data collected from physical examination are loosely coded with shorthand conventions and abbreviations known to health professionals and reflect the stereotypical examination process. The other narrative data are extremely difficult to standardize and significant problems can be associated with nonstandard abbreviations once they can have different meanings depending on the context in which they are used. Some attempts have been made to use conventional text notation as a form of summarization and complete phrases are used as loose standards. The enforcement to summarize heterogeneous conditions characterizing a simple concept about a patient is often unsuccessful [23].

Many data used in medicine are numerical. These include vital signs such as temperature or blood pressure, laboratory tests results, and some measurements taken during physical examination. This type of data is easier to formalize and some can be acquired and stored automatically. However, when numerical data are interpreted, the problem of precision and validity becomes important. In some fields of medicine, especially intensive care medicine, data are acquired in the form of continuous signals such as electrocardiogram or pulse oximetry wave. When these data are stored in medical records, a graphical tracing is frequently included. How this type of data is best managed in computer storage and clinical decision systems is an important challenge.

Visual images, acquired from machines or drawn by the health professionals to locate abnormalities or describe procedures, are an important type of data. Radiology images can be stored in electronic patient records using formalized compressing protocols. The possibility of acquiring and storing sketched images is an important challenge in the development of an electronic patient record due to its heterogeneity and lack of standardization.

Medical practice is medical decision-making [24]. Information exists to support decisions and actions such as procedures; if it fails to do this, it becomes irrelevant noise. The process of diagnosis is a probabilistic clinical reasoning based on three types of information: patient data, medical knowledge, and "directory" information (e.g., available surgical rooms at the hospital) [25]. This is why the diagnosis and the consequent clinical decision are an uncertain dynamic and an evolutive process. Although this type of information can be coded and structured, its storage in a patient record can be difficult to formalize due to this uncertainty, the existence of more than one diagnosis hypotheses, and the time-related evolution of the final diagnosis. Clinical procedures are easier to formalize and are usually represented with codes.

In general, health information can be divided into different groups. To each group, it is possible to associate a usual method to obtain such data and also the difficulty of formalization. Table 4.2 summarizes some of the characteristics of these groups.

4.2.4 Understanding What Is Recorded

Poor presentation of clinical data can also lead to poorly informed clinical practice, inappropriate repeated investigation, or unnecessary referrals, and wastes clinical time and other resources [3]. What humans understand is profoundly shaped by the manner in which data are presented, and by the way we react to different data presentations [4]. Thus, it is probably as important to structure the data in a way so that it can be best understood, as it is to ensure that the data are correct in the first place. The manner in which data are presented should take into consideration the current clinical context and anticipate the users' needs, thus creating an intelligent ambient [26].

TABLE 4.2

Groups of Health-Related Data

Type of Health Data	Method to Obtain Data	Formalization Difficulty	Type of Computer Data	Example of Standards
Narrative data	Ask the patient	Difficult to formalize	Free text	Some loose standards of conventional text notation
Physical examinations	Observation and measurements	The measurements are easier to formalize than the observations that are narrative data	Text and numeric data	LOINC, HL7, openEHR
Diagnosis	Reasoning	Some parts are difficult to formalize (e.g., uncertainty)	Numeric coded data	Several international codification standards such as SNOMED and ICD. Attempt to structure diagnosis using different classification branches
Procedures	Result of actions	Easy to formalize	Numeric coded data	Several international codification and classification standards on medical and surgical procedures as well as pharmacological therapy
Laboratory reports	Tests mainly done automatically	Easy to formalize	Numeric	LOINC
Images and biological signals	Measures made by machines	Easy to formalize	Numeric	DICOM

Note: DICOM, Digital Imaging and Communications in Medicine; HL7, Health Level Seven; ICD, International Classification of Diseases; LOINC, Logical Observation Identifiers Names and Codes; SNOMED, Systematized Nomenclature of Human and Veterinary Medicine.

In order to properly interpret EHR data, it is very important for the researcher to understand how the EHR IS works, that is, the software, hardware, people, and the processes. Thus, researchers using EHR data should have access to documentation describing database models, user forms, devices used in data collection, a description of the users, and the protocols used in data collection.

Moreover, to safely interpret health data from heterogeneous systems for research use, it is important to be able to share and communicate the meaning of data. Due to a potentially very large amount of data and necessary reasoning of the data, it is crucial that the semantics of the data are computer-interpretable. Such semantic concerns can generally be divided into data values, data structures, and terminology-related semantics.

To improve data accuracy, EHR should allow the association of a reliability measure to some data (e.g., "Suspect influenza"). Some authors have even proposed that associated to each value in a database, there should be a second value describing the reliability of the value [27]. The proposed categories are good, medium, poor, unknown, and not applicable.

4.3 Data Quality

With the development of informatics technology, medical databases tend to be more reliable. However, issues regarding DQ have become more relevant than ever as the utilization of these databases is increasing both in magnitude and importance. DQ is relative to each objective and can be defined as "fitness for use," that is, data can be considered of appropriate quality for one purpose but they may not hold sufficient quality for another situation [28]. This is especially true in medical databases; a medical database can be of quality for economic analyses but may be insufficient quality for a clinical study.

For data to have quality, Wyatt and Liu [29] stated that they should be accurate, complete, relevant, timely, sufficiently detailed, appropriately represented (e.g., consistently coded using a clinical coding system), and should retain sufficient contextual information to support decision-making. Other authors consider four dimensions of DQ [28]: accuracy, as the degree of correctness and precision with which real-world data are represented; completeness, as the degree to which all relevant data are recorded; consistency, as the degree to which data satisfy specified constraints and business rules; and timeliness, as the degree to which the recorded data are up-to-date [30, 31].

In another perspective, it is possible to analyze DQ concerning three roles about data: production, custodian, and consumer. Data producers are

those that generate data (e.g., medical, nursing, or administrative staff); data custodians are those that provide and manage computing resources for storing and processing data (e.g., database administrators and computer scientists); and data consumers are those who use data in medical care (e.g., physicians, researchers, and managers) [32, 33].

4.3.1 Problems in the Input of Data

Data in health records should be accurate, complete, and up-to-date, in order to be useful, not only for healthcare practice (its main purpose) but also for further research activities. Paper-based medical records (PBMRs) are still an important foundation of information for healthcare, and are often considered the gold standards for the evaluation of EHR systems, as they represent the closest contact with the actual event they report. The mauve reputation of PBMRs comes mostly from the fact that they require users to expend considerable time and effort to search for specific information, or to gather and obtain a general overview. Also, the input of information is performed by different persons, at different points in time, and is often done after the medical service has been administered, with main problems of poor handwriting, missing sheets, and imperfect documentation usually connected to the high workload of both physicians and nurses [34].

Over the past decades, a wide range of computer systems have been introduced to support clinical practice [35]. However, computerization does not necessarily help . For example, Soto et al. [36] reported in their study on documentation quality that, despite the presence of an electronic medical record designed to facilitate documentation, rates of documentation of some domains fell below desirable levels. EHRs have been weakened by both misconceptions on the record design, and shallow (or nonexistent) research on the design of user interfaces used to fill in and extract data. To help clinicians find data faster and with less effort, everyone designing and writing in records needs to understand how and why we search records and the design features that make searching easier. On one hand, good record design can double the speed at which a practiced reader extracts information from a document, whereas poor design introduces an upper limit on speed that cannot be overcome by training . On the other hand, the design of the user interface has a substantial influence on the usability of the system, and this plays an important role in the prevention of incomplete and incorrect data records [35]. Even considering perfect record and user interface designs, there is a high level of uncertainty in recorded data, some of it stemming from the fact that different users participate in data recording [36] and some stemming from the fact that data nowadays are being introduced in different record systems almost simultaneously or, at least, during the same patient's event, leading to record and event linkage problems [37].

4.3.2 Types of Errors

Arts et al. [38] reviewed the main conceptions about types of data errors, ranging from semantic specifications (e.g., interpretation, documentation, and coding errors) to underlying process definitions (e.g., systematic and random errors). Focus was given to the latter. Causes of systematic data errors include programming errors, unclear definitions for data items, or violation of the data collection protocol. Random data errors, for instance, can be caused by inaccurate data transcription and typing errors or illegible handwriting in the patient record [38]. They also conducted a simple case study to try to identify the main causes of data incompleteness and inaccuracy. They evaluated data recorded in two different intensive care units (ICUs), one using a PBMR and the other a patient data management system, both registering their own data into a central registry database at the National Institute for Health and Clinical Excellence coordinating center. The main sources of errors were searched for in the following three processes: local data recording, local data extraction, and central data transfer. From 20 randomly selected patients from each ICU, the overall evaluation resulted in the following: 2.0% inaccurate data (mostly for programming errors) and 6.0% incomplete data (mostly for poorly designed record, with missing variables) for the hospital using the automatic data collection; 4.6% inaccurate data (mostly for inaccurate transcription and/or calculations) and 5.0% incomplete data (almost evenly for the transcription from the PBMR and for programming errors in the data transfer process) for the hospital using manual data collection. Although this scenario is not one of the toughest ones regarding DQ, it revealed some interesting causes that need to be taken into account in EHR design. Tables 4.3 and 4.4 present the authors' overview of causes of insufficient DQ, in two different periods of a distributed EHR system implementation and deployment: registry setup and data collection.

We redirect the reader to the referred study for a thorough report on the causes of insufficient DQ. Nonetheless, we believe that the particular important causes that we should address in this chapter are poor interface design; lack of adherence to guidelines, protocols, and data definitions; and insufficient information on the collected data. A taxonomy of DQ problems (DQPs), organized by granularity level, is presented in Table 4.5 [39]. Next, we present a short definition for each DQP based on Oliveira et al. [39], and present some examples:

Missing values—a required attribute not filled, that is, the absence of value in a mandatory attribute (e.g., the gender of a patient is missing).

Syntax violation—attribute value violates the predefined syntax (e.g., birth_date is not in the correct date format).

Domain violation—attribute value violates the domain of valid values [e.g., length of stay (LOS) contains a negative value]. If the attribute

TABLE 4.3

Causes of Insufficient Data Quality during Setup and Organization of the Registry [38]

Problems at the Central Coordinating Center	**Type of Error**
Unclear/ambiguous data definitions	Systematic
Unclear data collection guidelines	Systematic
Poor case record form layout	Systematic/random
Poor interface design	Systematic/random
Data overload	Random
Programming errors	Systematic
Problems at the Local Sites	**Type of Error**
Illegible handwriting in data source	Random
Incompleteness of data source	Systematic
Unsuitable data format in source	Systematic
Data dictionary not available to data collectors	Systematic/random
Lack of motivation	Random
Frequent shift in personnel	Random
Programming errors	Systematic

TABLE 4.4

Causes of Insufficient Data Quality during Data Collection [38]

Problems at the Central Coordinating Center	**Type of Error**
No control over adherence to guidelines and data definitions	Systematic
Insufficient data checks	Systematic/Random
Problems at the Local Sites	**Type of Error**
Nonadherence to data definitions	Systematic
Nonadherence to guidelines	Systematic
Calculation errors	Systematic/Random
Typing errors	Random
Insufficient data checks at data entry	Systematic/Random
Transcription errors	Random
Incomplete transcription	Random
Confusing data corrections on case record form	Random

data type is string, this DQP can additionally be divided into the following:

- Overloaded attribute—attribute value partially violates the domain: a substring of it is valid, whereas the remaining substring is invalid (e.g., first_name contains all the names of the patient).
- Misspelling error—attribute value contains a misspelled error. A misspelling error can occur due to either typing errors or

TABLE 4.5

Data Quality Problems by Granularity Level [39]

Data Quality Problem	Attribute/Tuple			Single Relation	Multiple Relations	Multiple Sources
	Attribute	Column	Row			
Missing values	×					
Syntax violation	×					
Domain violation	×					
Overloaded attribute/invalid substring	×					
Misspelling error	×					
Ambiguous value	×					
Incorrect value	×					
Violation of business rule	×	×	×	×	×	×
Uniqueness violation		×				
Existence of synonyms		×			×	×
Violation of functional dependency				×		
Approximate duplicate tuples				×		×
Inconsistent duplicate tuples				×		×
Referential integrity violation					×	×
Incorrect reference					×	×
Heterogeneity of syntaxes					×	×
Heterogeneity of measure units					×	×
Heterogeneity of representation					×	×
Existence of homonyms						×

lack of knowledge of the correct spelling (e.g., prostate and prostrate).

– Ambiguous value—the attribute value is an abbreviation or acronym (e.g., BPD can mean bronchopulmonary dysplasia or borderline personality disorder).

Incorrect value—attribute contains a value which is not the correct one, but the domain of valid values is not violated (e.g., age is 56 instead of 59).

Violation of business rule—this problem can happen at all granularity levels, when a given business domain rule is violated (e.g., patient_name must have at least two words, but there is a tuple where this constraint is not respected).

Uniqueness violation—two (or more) tuples* have the same value in a unique value attribute (e.g., the same process_number for different patients).

Existence of synonyms—use of syntactically different values with the same meaning (within an attribute or among related attributes from multiple relations) (e.g., the use of gastric or stomach).

Violation of functional dependency—the value of a tuple violates an existing functional dependency among two or more attributes (e.g., in the same hospital, (dept_code = 40; dept_name = Gastroenterology) and (dept_code = 40; dept_name = Psychiatry)).

Approximate duplicate tuples—the same real-world entity is represented (equally or with minor differences) in more than one tuple [e.g., tuple department(Pneum, 8th floor, St. John Hospital) is an approximate duplicate of tuple department(Pneumology, eight floor, St. John Hospital)].

Inconsistent duplicate tuples—representation of the same real-world entity in more than one tuple but with inconsistencies between attribute values ([.g., address in duplicate tuples hospital(St. John Hospital, Great Ormond Street) and hospital(St. John Hospital, Saintfield Road) is inconsistent).

Referential integrity violation—a value in a foreign key attribute does not exist in the related relation as a primary key value (e.g., principal diagnosis code 434.91 is not present in the diagnosis relation).

Incorrect reference—the referential integrity is respected but the foreign key contains a value, which is not the correct one (e.g., principal diagnosis coded as 434.10 instead of 434.11; both codes exists in the diagnosis relation).

Heterogeneity of syntaxes—existence of different representation syntaxes in related attributes, within a data source or among data sources (e.g., attribute admission_date has syntax dd/mm/yyyy, but attribute discharge_date has syntax yyyy/mm/dd).

Heterogeneity of measure units—use of different measure units in related attributes, within a data source or among data sources (e.g., in different data sources, the representation of attribute temperature in different scales (Celsius/Fahrenheit).

* In the text of relational databases, tuple is formally defined as a finite function that maps field names to values in the relational model.

Heterogeneity of representation—use of different sets of values to code the same real-world property, within a data source or among data sources (e.g., in one source, gender can be represented with values 1, 2 and, in another source, with values M, F).

Existence of homonyms—use of syntactically equal values with different meanings, among related attributes from multiple data sources (e.g., ventilation has at least two different meanings, one referring to the biological phenomenon of respiration, and the other referring to the environmental flow of air).

4.3.3 Data Cleansing

Outliers, inconsistencies, and errors can be included in the process of data cleansing [40]. Data cleansing consists of the exploration of data for possible problems and making an effort to correct errors. This is an essential step in the process of knowledge discovery in databases (KDD) [41]. There are many issues related to data cleansing that researchers are attempting to tackle, such as dealing with missing data and determining record usability and erroneous data [42]. There is no general definition for data cleansing as it is closely related with the area where it is applied, for example, in KDD, data warehousing, or total DQ management [40].

The subgroup of DQPs associated with database outliers, inconsistencies, and errors can be divided into semantic, syntactic, and referential integrity problems [43]. The characterization and systematization of these problems is very important. An object that does not comply with the general behavior of data is called an outlier [41]. Outliers can happen due to mechanical faults, changes in system behavior, fraudulent behavior, human errors, or instrument errors or faults [44]. They are, normally, very different from, or inconsistent with, the remaining data. An outlier can be an error but can also result from the natural variability of data, and can hold important hidden information. There is no universal technique for the detection of outliers; various factors have to be considered [18]. Statistics and machine learning contribute with important different methodologies for their analysis [41, 45–47]. Usually, computer-based outlier analysis methods follow a statistical, a distance-based, or a deviation-based approach [41]. Algorithms should be selected when they are suitable for the distribution model, the type of attributes, the speed, the scalability, and other specific domain characteristics [44].

The manner in which these potential DQPs are managed is also very important. It is necessary to analyze and define what to do in each situation (e.g., to delete or label an error). To facilitate outlier analysis, a probability can be assigned to each case. Issues related to data preprocessing are also essential for the reduction of computational efforts to achieve a faster domain comprehension and a faster implementation of methods.

4.3.4 Missing Values

Missing values can represent system-missing (e.g., resulting from not selecting any option) or user-missing (e.g., resulting from selecting an option named "unknown"). System-missing should be avoided and replaced by user-missing whenever possible. The main problem regarding system-missing values arises at analysis time, namely, by trying to find out what it means when a null or a blank value is stored in the database. To illustrate this problem, let us consider an example form for the introduction of information about an allergy. Usual forms consider only a simple checkbox.

Allergy to penicillin: [ ]

What does the blank value mean? Is the patient not allergic? The physician does not know? The question is not applicable in the current setting? The value has not yet been introduced, but will probably be in following iterations?

Clearer forms consider a yes/no radio button to force the user to give an answer, even if it is the default answer.

Allergy to penicillin: ( )Yes (*)No

But even this is not enough to clarify things for the analyst. More complex forms should consider other hypotheses (N/A means Not Applicable):

Allergy to penicillin:

( )Yes ( )No (*)Unknown

( )Yes ( )No (*)Unknown ( )N/A ( )Yes ( )No ( )Unknown ( )N/A (*) Never entered

( )Doctor says Yes ( )Doctor says No

( )Patient says Yes ( )Patient says No

( )Unknown (*)Never entered ( )N/A

A similar approach has been suggested in a study on demographic surveillance systems [48]. Aiming to improve data reliability, a set of standard values was defined to be used consistently throughout the database to indicate the status of a particular data value. The following standard values (and their meanings) were proposed:

- "Never entered"—This is the default value for all data fields in a newly created record.
- "Not applicable"—Given the data in related fields or records, a value for this data field is not applicable.
- "Unknown"—The value is not known. Follow-up action yielded no better information or is not applicable.
- "To be configured"—This indicates a need to query the value as it appears on the input document and to take follow-up action.

"Out of range"—The value on the input document is out of range and could not be entered. Follow-up action yielded no better information or is not applicable.

The last two standard values ("To be configured" and "Out of range") are more closely related to the demographic surveillance systems study, and therefore more difficult to generalize to other uses.

Moreover, software data entry programs requiring the user to enter a value into a field, regardless of whether an entry would be applicable, may find that a value has been entered merely to satisfy the requirements of the program rather than to record valid data. Having options related to missing data helps to solve these questions.

4.3.5 Default Values

Default values are preassigned content for a data container (e.g., form entry or table field). It is a value that is used when no value is provided. To increase simplicity of use, computer user interfaces often use default values. Default values are normally the most common value, and they are assumed to be true unless the user specifies otherwise. Such default assumptions are not appropriate for the entry of clinical data, for two reasons [49]:

1. They may bias the data, because they may appear to be the "expected" or "normal" value.
2. The user may simply miss a question if defaults are taken.

The alternative is to make all data entry fields blank or initially set to "unknown," so that each data point in the database represents an explicit input by the patient. Palmblad and Tiplady [49] stress that responses should always result from an action by the user—defaults should not be taken as data.

4.3.6 Clinical Concepts

There are several sources of problems when designing user interfaces, especially in the healthcare domain. In the study of Hyeoneui et al. [50], problems were found on an ICU nursing flow sheet related to data item labels. Some labels (e.g., "Status" and "Condition") were found to be vaguely defined. They did not convey sufficient semantic information about the data they contain, making it difficult to retrieve and perform inference on the data.

Coded data offer a possible way to apply statistics to the collected data. The advantages are obvious. Nevertheless, coded data always imply a simplification of the reality. Structured form is a model of the reality, and as such they aim to transform a complex reality in a simplified version. Since free text does not impose so many restrictions, information loss tends to be much

lower. Also, when people are forced to select from a predetermined list of codes, they may discover that they cannot find the correct code and so they select a code that seems to be closest to, but does not truly represent, the real situation or observation [51].

For example, the accuracy of coded hospital information on a Patient Administration System of the Birmingham Women's Hospital was tested [52]. The accuracy of diagnosis, interventions, and diagnosis-intervention pairs on electronic records was low. The reported kappa (κ) agreement statistics were 0.39, 0.30, and 0.21, and the proportion of agreement was 69.69, 64.64, and 60.26, respectively [53]. It is important to note that these data had been previously used to measure the level of evidence-based healthcare [54]. The authors claim that a major source of error arose because the databases of maternities and surgical operations were not seamlessly linked to the Patient Administration System. This study concludes that a high degree of inaccuracy may exist in hospital electronic clinical data, and that researchers relying on hospital electronic data are advised to first check the level of accuracy.

In another example, the agreement on final diagnosis between two sources for pediatric emergency department (ED) visits in 19 U.S. institutions was recently studied [55]. Overall, 67% of diagnoses from the administrative and abstracted sources were within the same diagnosis group. Agreement varied by site, ranging from 54% to 77%. The authors concluded that the ED diagnoses retrieved from electronic administrative sources and manual chart review frequently disagreed. Agreement varied by institution and by diagnosis. Further work is recommended to improve the accuracy of diagnosis coding. Other studies have described misclassification issues in EHRs [56]. Coding problems increase when trying to integrate databases. Creating a common coding system from different sets of codes is complicated because it is probable that different processes for coding and different definitions were used. Below, four case studies related to coding issues are described.

Validation rule case study. In a central hospital discharges database, a simple validation tool, with domain rules defined by a specialist in medical codification, was implemented [57]. This tool periodically produces error reports and gives feedback to database administrators. With this feedback tool, and in only 7 months, critical errors decreased from 6% to 1% (from 173,795 to 27,218 cases). After that, critical errors continued to decrease and in 1 year were reduced to 0.13% (5.112 cases). This is a simple, yet very interesting, example on how simple procedures can clearly influence the quality of data and consequently the quality of any research using that data. Let us analyze an example of a validation rule (algorithm) for birth weight (see Section 1.5.6). Newborns are coded by having the principal International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) diagnosis code starting with V3. In these cases, the birth weight attribute (BIRTH_WGT) cannot be missing (Figure 4.2).

Influenza coding among hospitals case study. More complete examples of erroneous data can be found in the same data. ICD-9-CM has diagnostic

- Is the attribute filled?
  - No → principal diagnosis has a code starting with V3?
    - → Yes → output a critical error
    - → No → ok
  - Yes → principal diagnosis has a code starting with V3?
    - → Yes → is BIRTH_WGT value between 400 and 9000?
      - Yes → ok
      - No → output a critical error
    - → No → output a critical error

FIGURE 4.2
Example of validation rule for birth weight.

codes specific to influenza (487.0, 487.1, and 487.8) that can be easily retrieved from hospital discharge records. Although there have several successful ventures in the use of these codes [58], this specific codification is not uniformly done in different hospitals. In fact, using the previous national inpatient database, we can find substantial differences within different acute care hospitals (see Figure 4.3). The percentage of cases with influenza ranges from 1.77% to 0.00% of all hospitalizations in Portuguese hospitals. It is probable that most of the diagnoses of influenza were not introduced.

Ischemic stroke coding protocol case study. ICD-9-CM discharge data have also been used to identify patients with stroke for epidemiological, quality of care, and cost studies [59]. Nevertheless, for many years ischemic stroke (a poorly defined type of stroke) did not have a direct entry in the ICD-9-CM alphabetic index. In Portugal, due to an erroneous interpretation from an official entity, ischemic stroke was initially coded with 437.1 (other generalized ischemic cerebral-vascular disease), instead of the correct 436 code (acute but ill-defined cerebral-vascular disease). In October 2004, the Cooperating Parties and the Editorial Advisory Board for Coding Clinic for the ICD-9-CM clarified this situation and modified the classification system, pointing out the code 434.91 (cerebral artery occlusion unspecified with cerebral infarction) as the correct one for ischemic stroke. After a period of 2 to 3 years (time needed for the message to reach all medical coders in Portugal, because the use of the ICD-9-CM version is not up-to-date nor uniform), ischemic stroke coding started to be generally, and correctly, classified with 434.91. As shown in Figure 4.4, changes in the protocol for coding ischemic stroke clearly influenced ICD-9-CM discharge data: there is a clear reduction over years in episodes coded as 436 (not specified stroke) and a simultaneous increase in 434.91 (cerebral-vascular disease, including ischemic stroke).

Leukemia incidence decrease difficult to explain. Figure 4.5 shows the incidence (the number of new cases in a specified period) by year for leukemia,

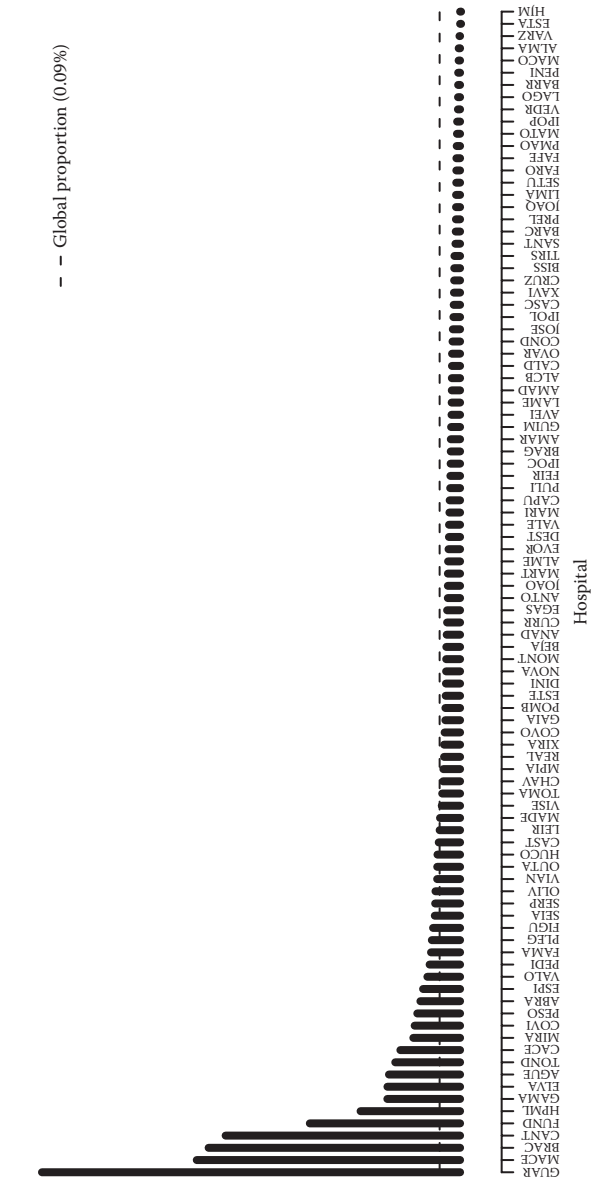

FIGURE 4.3
Influenza variation by hospital.

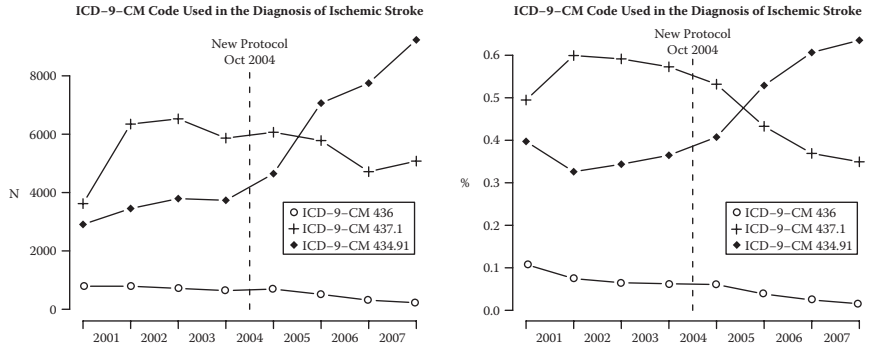

FIGURE 4.4
Evolution of ischemic cerebral-vascular disease coding.

calculated using the same national hospital discharge database. Medical specialists do not find a medical reason for such a decrease between years 1999 and 2001 (1183 to 770 cases). On the other hand, coding specialists suggest that perhaps a modification in coding policies could be the cause of this sudden decrease.

Unambiguous and consistent representation is the foundation of data reuse. Locally developed systems often fail to meet this requirement. There are methods to disambiguate concept representation [60]:

Top-down approach—identification of key concepts of a domain, which are then organized into a structure; then detailed concepts are filled into the structure.

Bottom-up approach—collection of all detailed concepts, which are then organized into a hierarchical structure.

In a 2008 study, Hyeoneui et al. [50] showed how these methods can be applied in a real-case scenario. A locally developed ICU nursing flow sheet was studied with the aim of extracting the conceptual model existing in the IS. Although a labor-intensive process, the disambiguation methods proved to be feasible in clarifying many ambiguous data representation in the local ICU nursing flow sheets.

4.3.7 Protocol of Data Collection

Pulling data elements out from the systems and environments that generate them also pulls them away from valuable information that gives them definition and structure. For example, for a single data element of blood pressure, its definition is fairly well understood. However, how one interprets blood pressure may differ with respect to: where it was measured (in the quiet confines

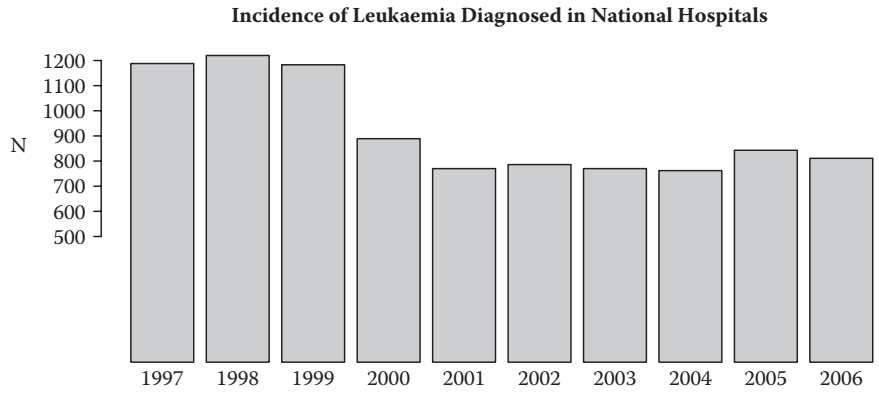

FIGURE 4.5
Incidence of leukemia in Portugal between 1997 and 2006, according to hospital coded data.

of a doctor's office or at the back of an ambulance); how it was measured; when it was measured (time and day); and who took the measurement.

Many questions may arise regarding data collection protocols:

Should it only store the data on blood pressure or should the protocol be also used to collect it?

Are values understandable without the protocol?

Can two temperatures, measured in different locations, represent the same variable?

Can two temperatures, measured with different devices, represent the same variable?

Savage [18] proposed that the following acquisition methods should be described in order to better understand the data:

- Business practices
- Measurement, observation, and assessment methods
- Recordkeeping practices
- Constraints and rules applied
- Changes in methods over time

In openEHR standard [61], special attention was given to the way health-related data are measured, so all the CARE ENTRY classes in the openEHR EHR Reference Model have a protocol section. This includes the OBSERVATION, EVALUATION, ACTION, and INSTRUCTION classes. The protocol section was added early in the design of openEHR based on research at the time demonstrating that computerization meant that a lot of details could be

added to the documentation that could be of use in the future. This section is used to record information that is not critical, but may add value, to the interpretation of a measurement. This often includes information about the manner in which something was measured, such as the device or location of a measurement (when the measurement value is applied to the body—such as temperature).

A different illustrative example shows how differences in protocol have an important impact on data interpretation. In Portugal, LOS is calculated by subtracting the admission date from the discharge date. Same-day stays are consequently coded as 0. Leave days are not subtracted. Even considering admission and discharge time, the process of calculating LOS is not always the some among different hospitals throughout the country. For same-day stays we, find, in some cases, LOS with a value of 0 and, in other cases, with a value 1. Other national administrative/clinical applications do not use the standard definition and include the leave day. Consequently, an episode where the patient leaves in the day after admission is assigned 2 days for LOS.

4.3.8 Date and Time Issues

One important piece of information regarding clinical activities is the moment when they occurred. In many circumstances, the accuracy of time data has profound medical, medicolegal, and research consequences (e.g., child birth, death, surgery, anesthetics, or resuscitation). Regarding data mining, the analysis of time data allows users establish the order of events and the time lapse between each event, and thereby allow event linkage when integrating different databases or doing process mining on log data.

One of the problems stems from the fact that there are several unsynchronized mechanisms used to tell the time. There is an old saying, "A man with a watch knows what time it is. A man with two watches is never sure." During resuscitations, multiple timepieces are used, and many events occur within a short period. Inaccuracies make it impossible to accurately reconstruct the order of events, which increases liability risks in the event of a lawsuit [62]. Standard bodies of the medical informatics field have already proposed solutions, such as Consistent Time Integration Profile, although they were mainly intended to synchronize logs, authenticate users, and digitally sign documents.

Births per minute case study. Another problem is related to people rounding off the minutes (or seconds) of clinical events. To test this premise, the authors of this chapter have measured the frequency of births grouped by the minute of birth (0 to 59) in a central Portuguese hospital. The database used had more 10,000 births registered. Some of the calculated frequencies were as follows: 9.15% for 0 minutes, 4.94% for 30 minutes, 1.8% for 12 minutes, and 0.44% for 51 minutes (see Figure 4.6). The top 12 most frequent minutes were all multiples of 5. The chart clearly shows the health professional tendency to input minutes that are in multiples of 5. Today, it seems a nurse

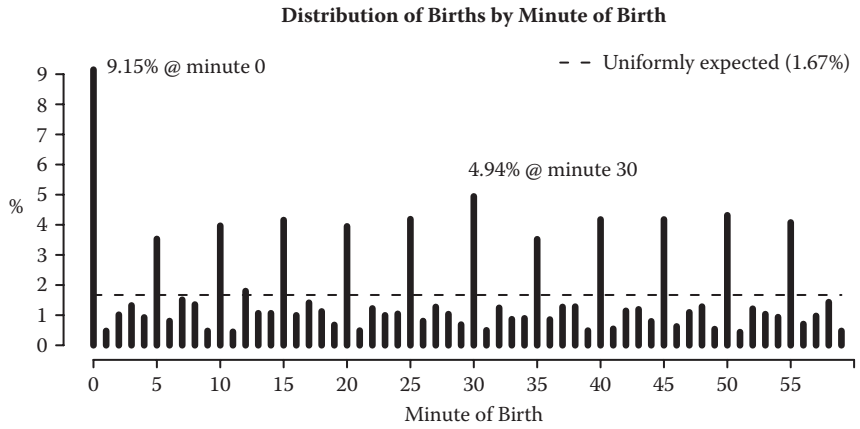

FIGURE 4.6
Frequency of births per minute.

with both a computer clock and a wall clock is never sure what time the birth really occurred. For several years now, nurses on the Perinatal Discussion List have questioned the dilemma of too many clocks ("Do you document the time from the wall clock, the electronic fetal monitor or the computerized medical record workstation clock?" "What if the times are discrepant?" "Where do you get standardized time?") [63].

A different study shows that 93% of cardiac arrest cases would contain a documentation error of 2 minutes (probably due to rounding off) or more and that 41% of cases would contain a documentation error of 5 minutes or more [64]. This value confirms the variation of critical timepiece settings in an urban emergency care system.

Kaye et al. [65] argue that the ability to use time intervals to evaluate resuscitation practice in the hospital is compromised by existing missing time data, negative calculated Utstein gold standard process intervals, unlikely intervals of 0 minute from arrest recognition to Advance Life Support interventions in units with cardiopulmonary resuscitation providers only, use of multiple timepieces for recording time data during the same event, and wide variation in coherence and precision of timepieces [65]. To detect such problems, the researcher should try to:

Understand the protocol/policy used by the EHR users to record time.

Find out how the different timepieces are synchronized.

Check if the devices and servers are similarly configured regarding time zones and daylight saving time.

Create algorithms to test the accuracy of data (e.g., the minute distribution of births should balanced).

In some clinical scenarios, there as been concern in solving this DQ issue. Ornato et al. [64] have documented that an attempted synchronization has cut a 2-minute documentation error rate in half and reduced the 5-minute documentation error rate by three-fourths. However, the error rates were predicted to return to baseline 4 months after the attempted synchronization [64]. They concluded that community synchronization of timepieces to an atomic clock can reduce the problem significantly, but the effects of a one-time attempted synchronization event are short-lived. In a more recent study, it was concluded that manually synchronizing timepieces to coordinated universal time improved accuracy for several weeks, but the feasibility of synchronizing all timepieces is undetermined [66].

Regarding the resuscitation clinical domain, Kaye et al. [65] argue that practitioners, researchers, and manufacturers of resuscitation equipment must come together to create a method to collect and document accurately essential resuscitation time elements.

Ideally, all devices and servers involved in the EHR should have their times synchronized (using Network Time Protocol). The researcher should also consider that in some countries, the official time changes twice a year (daylight saving time changes in summer and winter), complicating even further the analysis of data [67].

Births per day of month case study. It is also possible to find problems related to dates of events. The authors of this chapter have measured the frequency of births grouped by the day of the month (1 to 31) of a large database with all inpatient episodes of all Portuguese public hospitals. The values were adjusted according to the number of days. As shown in Figure 4.7, it is clear that the days of birth are also rounded to 1 or multiples of 5. A more detailed analysis allowed us to find higher differences among people born between 1918 and 1947 (range, 2.35–4.33%) and practically no differences among people born between 1978 and 2007 (range, 3.16–3.32%). This is probably related to the fact that people then used to register their children a few months after the actual birth, thereby increasing the possibility of rounding the registered day of birth.

4.4 Integration Issues of EHRs

Currently, people have more mobility and longer lives, and healthcare is more shared than ever. The need to integrate EHRs for healthcare delivery, management, or research is widely acknowledged. In its "Crossing the Quality Chasm" report, the Institute of Medicine [68] has documented the consequences of the absence of integration on the quality and costs of healthcare, and the need of "far greater than the current investments in information technology by most healthcare organizations." The main problem for the lack of

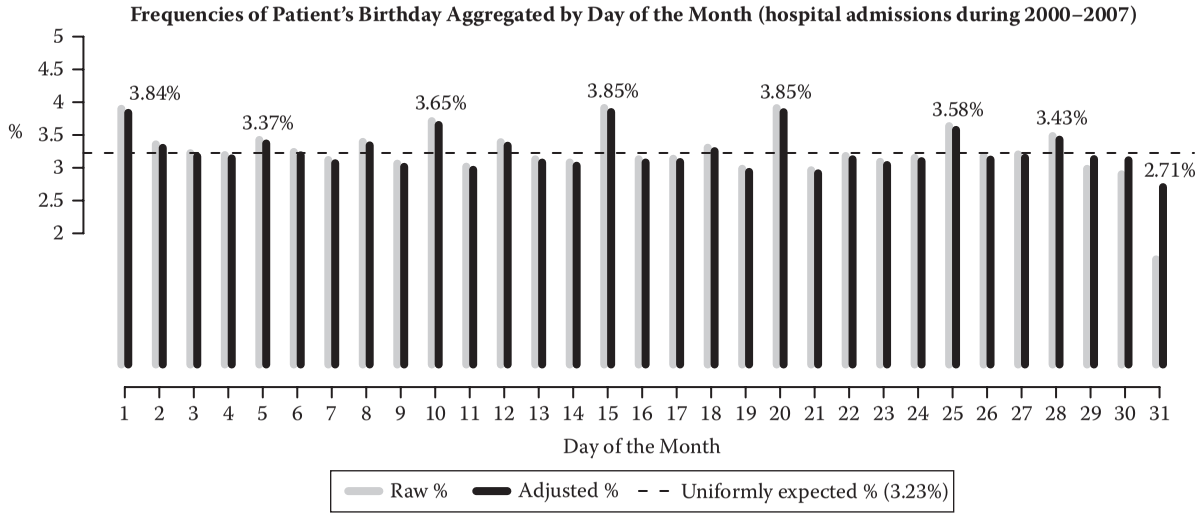

FIGURE 4.7
Frequency of births aggregated by day of the month. Data from 9 million records of all patients who had hospital admissions (inpatient episodes) in all Portuguese public hospitals between 2000 and 2007.

integration seems to be the poor incentive among healthcare institutions for overall integration [69]. Patients themselves, who have a strong incentive for integrated EHRs, are becoming the main drivers for the proliferation of the IS integration and Personal Health Records [70]. As patients become more aware and subsequently empowered in a patient-driven healthcare system, they will demand integrated EHRs as tools to help them manage their own health [71].

Clinical care increasingly requires healthcare professionals to access patient record information that may be distributed across multiple sites, held in a variety of paper and electronic formats, and represented as mixtures of narrative, structured, coded, and multimedia entries [72]. In hospitals, information technologies tend to combine different modules or subsystems, resulting in the coexistence of several IS aiming at a best-of-breed approach.

In a healthcare organization, processes are usually supported by several tools and there is a need to integrate these tools in order to achieve an integrated and seamless process [73]. Nevertheless, people will not willingly give up the stand-alone IS they use today because they fear data loss, loss of specific system functions customized to their needs, loss of control of their data (feeling that it represents their gold mine for research purposes), and they also have some pride about their own software implementation.

Integration of healthcare IS is essential to support shared care in hospitals, to provide proper care to mobile individuals, and to make regional healthcare systems more efficient. However, to integrate clinical IS in a manner that will improve communication and data use for healthcare delivery, research, and management, many different issues must be addressed [74–76]. Consistently combining data from heterogeneous sources takes substantial amounts of effort because the individual feeder systems usually differ in several aspects, such as functionality, presentation, terminology, data representation, and semantics [77]. It is still a challenge to make EHRs interoperable because good solutions to the preservation of clinical meaning across heterogeneous systems remain to be explored [72].

There are many standard bodies currently active in the formulation of international standards directly relating to EHRs. These are the International Organisation for Standardization (ISO), the European Committee for Standardization, and the Health Level Seven (HL7). Other organizations involved in the development of standards are the American Society for Testing and Materials, the Object Management Group, and the openEHR Foundation.

Over the years, different solutions to healthcare systems integration problems have been proposed and some have been applied. Many of these solutions coexist in today's healthcare settings and are influenced by technology innovation and changes in healthcare delivery. It should be noted that regarding standards in health informatics, there is the danger of confirming the ironic remark about the existence of so many of options that makes

our choice difficult. The fact that there are many solutions to health systems integration using different standards and data architectures may prove to be the greatest obstacle to semantic interoperability [78].

Lately, there has been an increasing number of publications describing projects, that integrate data from multiple IS [79]. This is in agreement with the assumption about the interest in improving the communication of health-related data to support person-centered healthcare. As the number of heterogeneous health IS grows, their integration becomes a priority. Moreover, we may be witnessing an increasing interest in regional integration among heterogeneous healthcare IS across different institutions, to bolster communication between the different stakeholders (primary and secondary care doctors, nurses, and patients). This is also supported by the increasing communication of referral letters. It is noteworthy that efforts are being expended toward integration in countries such as Germany, Greece, and Denmark, which are trying to implement nationwide healthcare integrated networks fed by heterogeneous IS.

Messaging technologies (in particular, HL7) are more used than middleware solutions (e.g., DICOM or CORBA). Web-based technologies (web services and web browsers) support most of the projects, indicating that new technologies are quickly adopted in healthcare institutions. Nevertheless, it is obvious that many distinct technological solutions coexist to integrate patient data.

The lowest semantics is about data values. There are several data types (e.g., text and numeric) that are supported universally on all major computing platforms. The use of standardized or common agreed data types (e.g., ISO data types) could further enhance the interoperability of low-level data semantics across systems. Based on that, certain ways of expressing value constraints for validation could enhance DQ and thus yield more reliable research.

On a higher level is the semantics of EHR information models based on which EHR systems are built. These EHR models provide the sense of structures in the EHR so that data entries can be locally grouped to meet clinical recording requirements, for example, outpatient encounter or inpatient admission screen forms. Occasionally, parts of the structures are also rendered on the screen in the form of headings, rubrics, and panels for easy navigation and usability. With modern EHRs, users have the choice to define their own particular structures to satisfy their clinical recording needs. Such mechanism is sometimes known as EHR templates [80–82]. Latest EHR interoperability technology even goes further to standardize the base EHR information models and allow these models to be further customized to meet volatile clinical requirements. ISO/EN 13606, openEHR archetypes/templates, and HL7 CDA templates are examples of these.

The use of standardized terminologies, medical vocabularies, and classifications in EHR provides the links between the data and externally defined concept models. Such links, sometimes known as terminology bindings,

are crucial to communicate the intended meaning of recorded data using concepts fully defined elsewhere. Such separation is necessary due to practical reasons. Concepts from terminology systems are universal and cover wide range from virus, bacteria, symptoms, to human anatomy, disease, and medical products. Authoring, managing, and maintaining these concepts and their relationships require access to experts from different domains of medicine, an effort that is only sustainable via international collaboration in order to achieve quality over long periods. This is exactly why international collaborations are now common (e.g., International Health Terminology Standards Development Organization). The rate of the change and the manner in which changes occur in these concept systems are quite different from those of the information models behind EHRs. The EHR information models are primarily designed for data recording to support care. Because of the changes in care protocols and processes, and the need for supporting *ad hoc* documentation, these EHR models are more volatile than the terminology systems and need to be close to where recording occurs.

4.4.1 Differences in Users

The users of EHRs are not all the same. They are different in terms of their background (e.g., medical doctors, nurses, radiology and laboratory technicians, and even patients), professional experience (e.g., young doctors doing their internship vs. senior specialists), or their computer experience (e.g., users typing long texts very fast vs. users having difficulties navigating through the IS). Physicians themselves enter the majority of data, either by populating specific fields or by keyboard entry of free text. Moreover, the exact time of entering data by physicians is also an issue, as most of the time this occurs immediately upon completion of the office visit; occasionally, this occurs during the visit in the presence of the patient, and sometimes at the end of a clinical session [36]. There are even cultural differences, which cannot be modeled by prior knowledge, that introduce high levels of uncertainty in the recorded data and the corresponding DQ.

A study that tried to unveil associations among users' clinical experience and DQ has been conducted by Soto et al. [36] in the ambulatory care setting. The study included primary care physicians (internists and pediatricians) and explored differences in the quality of medical records documentation according to several dimensions, including gender (both for physicians and patients), years since medical school, teaching status, and practice site (inside or outside the city). The objective was to measure the documentation of the patient's smoking history, drug allergies, medications, screening guidelines, and immunizations. Study results are paradigmatic of the type of (unexpected) differences that could rise in similar scenarios. Pediatricians and internists mainly differed in their patterns of documenting smoking status, which was expected as smoking status may be considered a more appropriate measure of quality for internal medicine than for pediatrics. However,

although pediatricians were more likely to document smoking status for their older patients, internists were less likely to document smoking status with increasing age of the patient, suggesting biased focus of smoking documentation on adolescents and young adults. The study also revealed differences in documentation regarding gender and experience. A sample of some of the probably less expected results for the internists group include: female internists were more likely than male internists to document smoking history, but less likely to document drug allergies; with increasing number of years since completing medical school, internists were less likely to document drug allergies and immunizations, whereas the increase in clinical time was associated with better documentation for smoking history. Overall, if some of the differences were somehow expected due to experience and clinical practice, gender had a priori unpredictable effects on the documentation quality. The authors concluded that no consistent pattern of correlates of medical record documentation quality emerged from the study, which might reinforce our view that there is a high randomness factor associated with differences among users.

Differences between users case study. A different setting arises when two or more users record a single event, especially if they belong to different professional groups. Consider the following example where both the emergency team and the firemen record the time at which an emergency team arrives at the event scene assigned to the event (in this scenario, the fire department is responsible for patient transportation). We gathered the time of arrival of the emergency team recorded by themselves and by the firefighters for 235 different events. The reader should recall that there is no "gold standard" in this problem; we can only make relative comparisons, and without any quality assessment of each group's record. Most of the time (60%), firemen recorded later times than those recorded by the emergency team, whereas only in 20.4% of cases did both teams agree on the arrival time. Figure 4.8 shows the distribution of the differences between recording times in the two groups, for all events. One event was recorded with a 54-minute (negative) difference between the two groups, and as a difference of more than 30 minutes seems more like recording error than a disagreement, we considered it an outlier and discarded this record. The remaining 234 records follow a symmetric distribution, with mean = 2.21 and standard deviation = 5.028 minutes of difference between the two record types. Assuming this sample is representative of the population of events occurring in similar scenarios, and that the mean is significantly above zero (99% confidence interval, 1.36–3.06), we could conclude that firemen tend to record later times than the emergency team. In this scenario, none of the teams has access to a "correct" time, each one using their own clock and way of registration (at the time or retrospectively), and yet the emergency team records arrivals as having occurred 2 minutes (median) earlier than firemen do. Moreover, the distribution of absolute differences when the emergency team reports later arrival times than firemen is also different from the distribution of events when firemen report

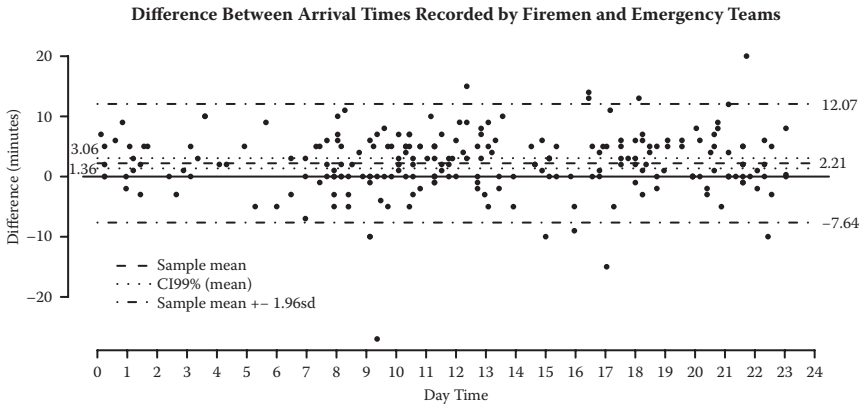

FIGURE 4.8
Differences between arrival times recorded by firemen and emergency teams.

later times (Mann-Whitney test, $p = 0.002$), with median values of 3 and 5 minutes (either by rounding or truncating the values), respectively, reinforcing the differences among the two groups of users. However, an important feature could be the basis of these differences: emergency teams have always registered event time in multiples of 5 minutes, whereas firemen have done it only around 45% of the times (nonetheless above the uniformly expected frequency of 20%).

4.4.2 Record and Event Linkage

One of the main challenges of health IS or networks is to be able to gather the different parts of the medical record of a patient without any risk of mixing them with those of another patient [83, 84]. Erroneous patient identification also has an impact on research and on hospital charging, as subsidiary partners refuse to pay for misidentified medical procedures. Record linkage between different IS is a risk but also an opportunity, because cross-checking between integrated distributed systems may be used to guarantee global patient DQ [85].

Patient identification errors case study. Cruz-Correia et al. [85] have studied the frequency of patients identification errors on clinical reports from four departmental IS and the hospital administrative database in a central Portuguese hospital. Table 4.6 presents the number of identification errors found as a new algorithm to cross-check data in different IS was being introduced. Furthermore, the assessment of the correctness of collected patient data is a difficult process even when we are familiar with the system under which it was collected [16].

Generally, record linkage refers to the task of finding entries or records that refer to the same entity (e.g., patient) in two or more files or databases. It

TABLE 4.6

Frequency of Patient's Identification Errors on Clinical Reports from Four Departmental Information Systems and the Hospital Administrative Database between July and December 2005

Information System	Total	Jul	Aug	Sep	Oct	Nov	Dec
System A	374	102	219	10	26	12	5
System B	44	12	7	5	5	11	4
System C	2					1	1
System D	1				1		
Hospital Administrative System	2	2					
Total errors found	423	116	226	15	32	24	10
Total reports checked	391.258	62.455	61.810	66.737	67.267	67.680	65.309

is an appropriate technique when the user needs to join data that are spread over more than one database. Record linkage is a useful tool when performing data mining. One typical use involves joining records of persons based on name when no national identification number or similar information is recorded in the data. The term record linkage was initially coined in 1946 by Dunn [86], who used it to designate the linking of various records of a person's life. As an example, record linkage could be used in mortality data sets in a cohort study, to determine who has (or has not) died [87]. The methods used in record linkage can be deterministic or probabilistic.

They are deterministic when they are based on defined identifiers for each individual that are usually assigned centrally and used in any records that are kept for that individual. It can be undertaken whenever there is a unique identifier, such as a personal identification number. A critique of the deterministic match rules is that they do not adequately reflect the uncertainty that may exist for some potential links [88].

They are probabilistic when they are based on combinations of nonunique characteristics of each individual, such as name, date of birth, or gender. It uses probabilities to determine whether a pair of records refers to the same individual. Patterns of agreement and disagreement between identifying characteristics are translated into quantitative scores, which are then used to predict whether the two records should be linked [89]. Statistics are calculated from the agreement of fields on matching and differing records to determine weights on each field. During execution, the agreement or disagreement weight for each field is added to obtain a combined score representing the probability that the records refer to the same entity. Often, there is one threshold above which a pair is considered a match, and another threshold below which it is considered not to be a match. Between the two thresholds, a pair is considered to be "possibly a match," and dealt with accordingly (e.g., human reviewed, linked, or not linked, depending on the application).

Regardless of the record linkage technique used, data normalization is always very important. Heterogeneity can be found in the formats of dates, people names, organizations, or department names, etc. By normalizing these data into a common format and using comparison techniques that handle additional variation, a much higher consistency can be achieved, resulting in higher accuracy in any record linkage technique.

An extension of patient record linkage based on health-related events (event linkage) is currently being studied [90]. This technique uses some demographic data in conjunction with event dates to match events in two data sets providing a possible method for linking related events.

4.5 Discussion

Imperfect data have a strong negative effect on the quality of knowledge discovery results. In this chapter, we focus on DQ issues that mainly depend on confounding causes, not necessarily visible (or even imaginable) to the data analyst. The main "take-away message" we intend to pass is that the data analyst should also consider expert knowledge about the overall setting, and about the data themselves—how it was collected, processed, and analyzed.

The data recorded on EHRs is the result of the processes of healthcare delivery. For the data to be fully understood, it is essential for the researcher to know such processes. The authors of this chapter argue that blindly analyzing EHR data without understanding the data collection process will probably lead to erroneous conclusions. Unfortunately, proper documentation on the data collection process is rarely forthcoming, and so the responsibility is left to the researcher to perform this task.

Regarding DQ, many different issues must be tackled with in EHR data (e.g., missing values, erroneous clinical coding, time synchronization). Currently, it is still very difficult to guarantee that EHR data are accurate, complete, relevant, timely, sufficiently detailed, and appropriately represented, and that they retain sufficient contextual information to support decision-making.

The integration of different EHRs further increases the difficulty of performing information extraction from EHRs. Due to the development of both EHR models and terminology systems, there is a growing overlapping of semantics that could be possibly represented using one technology over another. The best approach to perform terminology bindings between common EHR models and standardized terminology systems is still under intensive research, and the topic is probably beyond the scope of this chapter. It is nonetheless important to recognize such a need, the ongoing efforts, and what has already been established as recommendations in the field.

Once the important role of reference terminologies (e.g., SNOMED Clinical Terms), common concept models (EN-13940: CONTsys), and EHR interoperability standards (e.g., EN/ISO-13606, openEHR) have been fully recognized, it will be possible to share detailed and machine-interpretable care plans [91] not only to support continuity of care, but also to facilitate clinical research. In such scenario, care plans are instantiated from commonly agreed and/or evidence-based clinical guidelines, and can be communicated and understood by different EHR systems to provide guideline-based decision support to care providers across different organizations over potentially long periods. Because the care plans are based on guidelines and fully computerized [92], it would be feasible to check the guideline compliance of the care provided. Noncompliance treatment would either be followed up as quality issues or serve as input into clinical research for discovery of new knowledge.

Research on as well as other uses of EHRs are not likely reach their full potential before reference terminologies, common concept models, and EHR interoperability standards have been widely adopted. The challenge is still considerable, but with the recent EHR R&D projects [93] taking place in the academia, standardization bodies, and industry, what was deemed "an impossible task before is now a very difficult one."

In conclusion, although much research is still needed to improve DQ in some areas (e.g., semantic interoperability), there are already simple techniques available that should be made mandatory in EHR implementation to improve the quality of research and reduce misconceptions (e.g., time synchronization, proper database documentation, version control of data, and registration of all protocol changes). Meanwhile, researchers analyzing EHR data should be very careful with the interpretation on the retrieved results.

4.6 Related Work and Further Readings

Data anomalies can take a number of different forms, each with a different range of analytical consequences [94], focusing on outliers (definition of which is several times subjective), missing data, misalignments (strongly connected with time-related data problems), and unexpected structure (mostly appearing as multivariate outliers). Most research topics dealing with imperfect records consider statistical approaches to detect and correct these imperfections [95]. The impact of imperfect data on healthcare domains is enhanced by the fact that in many medical studies, particularly (prospective) studies of ongoing events, the patient group whose health is most threatened is represented by rather small numbers of subjects [96]. In a seminal data mining book, Breiman et al. [84] presented several examples of medical studies, where the use of decision trees proved to be helpful, but not without inspecting DQ. For example, when studying the diagnosis of heart

attacks, 3.6% of cases had to be excluded due to data incompleteness. Most of the works actually rely on this a priori analysis of anomalies to assure the quality of results. However, although all studies presented in the book reported missing data in the included cases, its impact on the results was not clearly discussed. Moreover, even if the protocol is followed "by the book," and no significant anomaly exists, there is also uncertainty in the data (e.g., if a sensor reads 100, most of times the real value is around 100—it could be 99 or 101), which should be taken into account.

A "good" data analysis result should be insensitive to small changes in either the methods or the data sets on which the analysis is based [94]. To produce robust and reliable results from such uncertain data sets, care has to be taken regarding the inferred conclusions. This idea is the mote for the *generalized sensitivity analysis* metaheuristic. A good introductory presentation on this procedure is presented by Pearson (Chapter 6) [94]. Basically, research has followed the path of a group of approaches that generate perturbations of initial learning set to assess the reliability of final models [97]. The bootstrap method [98] is a general tool for assessing statistical accuracy [99]. The main idea is to randomly draw data sets with replacement from the original data set, each of which has the same size as the original, and perform the analysis on all samples, and examine the fit of our model over the replications. *Bagging* [100] and *boosting* [101] are well known and possibly the most popular derivatives in the knowledge discovery field. They have been shown to improve generalization performance compared to individual models. Although *bagging* works by learning different models in different regions of the input space (by sampling original data set using specific bootstrap parameters), *boosting* focuses on those regions that are not so well covered by the learned model. These techniques perturb the entire learning data set that is fed to individual models, thus operating by creating different learning models [97].

Another related field deals with change mining, in the sense that data are not static and changes might occur and should be mined. When data are collected for long periods, the assumption that the process that is producing the examples is static does not hold. As seen in some of the examples we present in this chapter, such as the results on leukemia incidence and ischemic stroke coding, time matters. Stream mining and concept drift detection have been widely studied to implement such detection mechanisms [102]. In stream mining, data are processed as an open-ended continuous flow, where no (or little) storage is available [19, 103], so models should evolve with time. On one hand, mechanisms are necessary to detect changes in the underlying process producing the data stream [104]. On the other hand, change mining aims at identifying changes in an evolving domain by analyzing how models and patterns change [105]. This last subject is a hot topic in current data mining research, and their benefits to knowledge extraction from EHRs are clear, as the data in these changes are usually time-dependent. Further work should be considered in this field.

Acknowledgments

The authors would like to acknowledge the support given by the research project HR-QoD (outliers, inconsistencies, and errors) in hospital inpatient databases: methods and implications for data modeling, cleansing, and analysis (project PTDC/SAU-ESA/75660/2006). Also, the work of Pedro P. Rodrigues is supported by the Portuguese Foundation for Science and Technology (FCT) under PhD Grant SFRH/BD/29219/2006 and FCTs Plurianual financial support attributed to Laboratório de Inteligência Artificial e Apoio à Decisão. Finally, some of Rong Chen's work is part of NovaMedTech funded by Nutek, the Swedish Agency for Economic and Regional Growth, and the European Union Structural Funds.

References

1. Komaroff, A. L. 1979. The variability and inaccuracy of medical data. *Proceedings of the IEEE* 67(9):1196–1207.
2. Hogan, W. R., and Wagner, M. M. 1997. Accuracy of data in computer-based patient records. *Journal of the American Medical Informatics Association* 4(5):342–355.
3. Wyatt, J. C., and Wright, P. 1998. Design should help use of patients' data. *Lancet (British edition)* 352(9137):1375–1378.
4. Coiera, E. 2003. *Guide to Health Informatics*. London: Arnold London.
5. Barnett, O. 1990. Computers in medicine. *JAMA* 263(19):2631.
6. Richart, R. H. 1970. Evaluation of a medical data system. *Computers and Biomedical Research* 3(5):415.
7. Audit Commission. 1995. For your information: a study of information management and systems in the acute hospital.
8. Mamlin, J. J., and Baker, D. H. 1973. Combined time-motion and work sampling study in a general medicine clinic. *Medical Care* 11:449–456.
9. Korpman, R. A., and Lincoln. T. L. 1998. The computer-stored medical record: for whom? *Journal of the American Medical Informatics Association* 259:3454–3456.
10. Wyatt, J. C. 1994. Clinical data systems: Part 1. Data and medical records. *The Lancet* 344:1543–1547.
11. Dick, R. S., and Steen, E. B., eds. 1977. *The Computer-based Patient Record: An Essential Technology for HealthCare*. Washington, D. C.: National Academy Press.
12. Nygren, E., Wyatt, J. C., and Wright, P. 1998. Helping clinicians to find data and avoid delays. *The Lancet*, 352:1462–1466.
13. Hammond, K.W., Helbig, S. T., Benson, C. C., and Brathwaite-Sketoe, B. M. 2003. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. In *AMIA Annual Symposium Proceedings*, pp. 269–73.

14. Hohnloser, J. H., Fischer, M. R., Konig, A., and Emmerich, B. 1994. Data quality in computerized patient records. Analysis of a haematology biopsy report database. *International Journal of Clinical Monitoring and computing* 11(4):233–240.
15. Weir, C. R., Hurdle, J. F., Felgar, M. A., Hoffman, J. M., Roth, B., and Nebeker, J. R. 2003. Direct text entry in electronic progress notes—an evaluation of input errors. *Methods of Information in Medicine* 42(1):61–67.
16. Berner, E., and Moss, J. 2005. Informatics challenges for the impending patient information explosion. *Journal of the American Medical Informatics Association* 12(6):614–617.
17. Hogan, W. R., and Wagner, M. M. 1997. Accuracy of data in computer-based patient records. *Journal of the American Medical Informatics Association* 4(5):342–355.
18. Savage, A. M. 1999. Framework for characterizing data and identifying anomalies in health care databases. In *Proceedings of the AMIA Symposium*, p. 374. American Medical Informatics Association.
19. Muthukrishnan, S. 2005. *Data streams: Algorithms and Applications*. New York, NY: Now Publishers Inc.
20. Johnson, S. B. 1996. Generic data modeling for clinical repositories. *Journal of the American Medical Informatics Association* 3(5):328–339.
21. Van Ginneken, A. M., Stam, H., and Duisterhout, J. S. 1994. A powerful macro-model for the computer patient record. In *Proceedings of the Annual Symposium on Computer Application in Medical Care,* p. 496. American Medical Informatics Association.
22. Los, R. K. 2006. Supporting Uniform Representation of Data. PhD thesis, Department of Medical Informatics, Erasmus Medical Center, Rotterdam, the Netherlands.
23. Shortliffe, E. H., and Cimino, J. J. 2006. *Biomedical Informatics—Computer Applications in Health Care and Biomedicine*, 3rd edn. New York, NY: Springer.
24. Shortliffe, E. H., Perreault, L. E., Wiederhold, G., and Fagan, L. M. 1990. *Medical Informatics: Computer Applications in Health Care*. Boston, MA: Addison-Wesley Longman Publishing Co.
25. Wyatt, J. C., and Sullivan, F. 2005. *ABC of Health Informatics*. Blackwell Publishing, Malden, MA: BMJ Books.
26. Riva, G. 2003. Ambient intelligence in health care. *Cyberpsychology & Behavior* 6(3):295–300.
27. Anfindsen, O. J. 2000. Database management system and method for combining meta-data of varying degrees of reliability. US Patent 6,044,370.
28. Tayi, G. K., and Ballou, D. P. 1998. Examining data quality. *Communications of the ACM* 41(2):54–57.
29. Wyatt, J. C., and Liu, J. L. Y. 2002. Basic concepts in medical informatics. *British Medical Journal* 56(11):808–812.
30. Wang, R. Y. 1998. Total data quality. *Communications of the ACM* 41(2):58–65.
31. Gertz, M., Ozsu, T., Saake, G., and Sattler, K. 2003. Data quality on the Web. In *Dagstuhl Seminar*, Dagstuhl, Germany.
32. Strong, D. M. Lee, Y. W., and Wang, R. Y. 1997. Data quality in context. *Communications of the ACM* 40(5):103–110.
33. Orr, K. 1998. Data quality and systems theory. *Communications of the ACM* 41(2):66–71

34. Pourasghar, F., Malekafzali, H., Kazemi, A., Ellenius, J., and Fors, U. 2008. What they fill in today, may not be useful tomorrow: lessons learned from studying medical records at the women's hospital in Tabriz, Iran. *BMC Public Health* 8(1):139.
35. Jaspers, M. W., Knaup, P., and Schmidt, D. 2006. The computerized patient record: where do we stand. *Methods of Information in Med*, 45(Suppl 1):29–39.
36. Soto, C. M., Kleinman, K. P., and Simon, S. R. 2002. Quality and correlates of medical record documentation in the ambulatory care setting. *BMC Health Services Research* 2(1):22.
37. Roberts, C. L., Algert, C. S., and Ford, J. B. 2007. Methods for dealing with discrepant records in linked population health datasets: a cross-sectional study. *BMC Health Services Research* 7:12.
38. Arts, D. G. T., de Keizer, N. F., and Scheffer, G. J. 2002. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association* 9(6):600–611.
39. Oliveira, P., Rodrigues, F., and Henriques, P. 2005. A formal definition of data quality problems. In *IQ*, F. Naumann, M. Gertz, and S. Madnick, eds. Cambridge, MA: MIT Press.
40. Maletic, J. I., and Marcus, A. 2000. Data cleansing: beyond integrity analysis. In *Proceedings of the Conference on Information Quality,* pp. 200–209.
41. Kamber, M., and Han, J. 2001. *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.
42. Koh, H. C., and Tan, G. 2005. Data mining applications in healthcare. *Journal of Healthcare Information Management* 19(2):64–72.
43. Na, K. S., Baik, D. K., and Kim, P. K. 2001. A practical approach for modeling the quality of multimedia data. In *Proceedings of the 9th ACM international conference on Multimedia*, pp. 516–518. New York, NY: ACM Press.
44. Hodge, V., and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review* 22(2):85–126.
45. Lee, A. H., Xiao, J., Vemuri, S. R., and Zhao, Y. 1998. A discordancy test approach to identify outliers of length of hospital stay. *Statistics in Medicine*, 17(19):2199–2206.
46. Podgorelec, V., Hericko, M., and Rozman, I. 2005. Improving mining of medical data by outliers prediction. In *18th IEEE Symposium on Computer-Based Medical Systems, 2005. Proceedings*, pp. 91–96.
47. Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of data*, pp. 427–438. New York, NY: ACM Press.
48. Network, I. 2002. Population and health in developing countries: volume 1. *Population, Health, and Survival at INDEPTH sites*. IDRC, Ottawa, ON, CA.
49. Palmblad, M., and Tiplady, B. 2004. Electronic diaries and questionnaires: designing user interfaces that are easy for all patients to use. *Quality of Life Research* 13(7):1199–1207.
50. Hyeoneui, K., Harris, M. R., Savova, G. K., and Chute, C. G. 2008. The first step toward data reuse: disambiguating concept representation of the locally developed ICU nursing flowsheets. *Computers, Informatics, Nursing* 26(5):282.
51. Connell, F. A., Diehr, P., and Hart, L. G. 1987. The use of large data bases in health care studies. *Annual Review of Public Health* 8(1):51–74.

52. Ola, B., Khan, K. S., Gaynor, A. M., and Bowcock, M. E. 2001. Information derived from hospital coded data is inaccurate: the Birmingham Women's Hospital experience. *Journal of Obstetrics and Gynaecology* 21(2):112–113.
53. Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37.
54. Ellis, J., Mulligan, I., Rowe, J., and Sackett, D. L. 1995. Inpatient general medicine is evidence based. A-team, Nuffield Department of Clinical Medicine. *Lancet* 346(8972):407.
55. Gorelick, M. H., Knight, S., Alessandrini, E. A., Stanley, R. M., Chamberlain, J. M., Kuppermann, N., and Alpern, E. R. 2007. Lack of agreement in pediatric emergency department discharge diagnoses from clinical and administrative data sources. *Academic Emergency Medicine* 14(7):646–652.
56. Icen, M., Crowson, C. S., McEvoy, M. T., Gabriel, S. E., and Kremers, H. M. 2008. Potential misclassification of patients with psoriasis in electronic databases. *Journal of the American Academy of Dermatology* 59(6):981–985.
57. Silva-Costa, T., Freitas, A., Jácome, J., Lopes, F., and Costa-Pereira, A. 2007. A eficicia de uma ferramenta de validao na melhoria da qualidade de dados hospitalares. In *CISTI 2007—2 Conferncia Ibrica de Sistemas e Tecnologias de Informao*, Portugal, June.
58. Keren, R., Wheeler, A., Coffin, S. E., Zaoutis, T., Hodinka, R., and Heydon, K. 2006. ICD-9 codes for identifying influenza hospitalizations in children. *Emerging Infectious Disease* 12(10):1603–1604.
59. Goldstein, L. B. 1998. Accuracy of ICD-9-CM coding for the identification of patients with acute ischemic stroke: effect of modifier codes. *Stroke* 29(8):1602–1604.
60. Rassinoux, A. M., Miller, R. A., Baud, R. H., and Scherrer, J. R. Modeling concepts in medicine for medical language understanding. *Methods of Information in Medicine* 37(4–5):361–372.
61. Kalra, D., Beale, T., and Heard, S. 2005. The openEHR Foundation. *Studies in Health Technology and Informatics* 115:153.
62. Synchronize timepieces in your trauma room. 2000. *ED Management*, 12(2):23–24.
63. McCartney, P. R. 2003. Synchronizing with standard time and atomic clocks. *MCN: The American Journal of Maternal Child Nursing* 28(1):51.
64. Ornato, J. P., Doctor, M. L., Harbour, L. F., Peberdy, M. A., Overton, J., Racht, E. M., Zauhar, W. G., Smith, A. P., and Ryan, K. A. 1998. Synchronization of timepieces to the atomic clock in an urban emergency medical services system. *Annals of Emergency Medicine* 31(4):483–487.
65. Kaye, W., Mancini, M. E., and Truitt, T. L. 2005. When minutes count: the fallacy of accurate time documentation during in-hospital resuscitation. *Resuscitation* 65(3):285–290.
66. Ferguson, E. A., Bayer, C. R., Fronzeo, S., Tuckerman, C., Hutchins, L., Roberts, K., Verger, J., Nadkarni, V., and Lin, R. 2005. Time out! Is timepiece variability a factor in critical care? *American Journal of Critical Care* 14(2):113.
67. Neumann, P. 1995. *Computer-Related Risks*. New York, NY: ACM Press.
68. Institute of Medicine. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, D. C.: National Academy Press.
69. Herzlinger, R. E. 2004. *Consumer-Driven Health Care: Implications for Providers, Payers, and Policy-Makers*. San Francisco, CA: Jossey-Bass.

70. MacStravic, S. 2004. What good is an EMR without a PHR? *HealthLeaders*, September 3.
71. Kukafka, R., and Morrison, F. 2006. Patients' needs. In *Aspects of Electronic Health Record Systems*, H. P. Lehmann et al., eds. pp. 47–64. Calgary: Springer.
72. Kalra, D. 2006. Electronic health record standards. *Methods of Information in Medicine* 45(1):136–144.
73. Land, R., and Crnkovic, I. 2003. Software systems integration and architectural analysis—a case study. In *Proceedings of the International Conference on Software Maintenance, ICSM 2003*, pp. 338–347.
74. Heathfield, H., Pitty, D., and Hanka, R. 1998. Evaluating information technology in health care: barriers and challenges. *British Medical Journal* 316:1959–1961.
75. Berg, M. 2001. Implementing information systems in health care organizations: myths and challenges. *International Journal of Medical Informatics* 64(2–3):143–156.
76. Littlejohns, P., Wyatt, J. C., and Garvican, L. 2003. Evaluating computerised health information systems: hard lessons still to be learnt. *British Medical Journal* 326:860–863.
77. Lenz, R., and Kuhn, K. A. 2002. Integration of heterogeneous and autonomous systems in hospitals. *Business Briefing: Data management & Storage Technology*.
78. Ferranti, J., Musser, C., Kawamoto, K., and Hammon, E. 2006. The clinical document architecture and the continuity of care record: a critical analysis. *Journal of the American Medical Informatics Association* 13(3):245–252.
79. Cruz-Correia, R. J., Vieira-Marques, P., Ferreira, A., Almeida, F., Wyatt, J. C., and Costa-Pereira, A. 2007. Reviewing the integration of patient data: how systems are evolving in practice to meet patient needs. *BMC Medical Informatics and Decision Making* 7(1):14.
80. Chen, R., Enberg, G., and Klein, G. O. 2007. Julius—a template based supplementary electronic health record system. *BMC Medical Informatics and Decision Making* 7(1):10.
81. Los, R. K., van Ginneken, A. M., and van der Lei, J. 2005. OpenSDE: a strategy for expressive and flexible structured data entry. *International Journal of Medical Informatics* 74(6):481–490.
82. Hoya, D., Hardikerb, N. R., McNicollc, I. T., Westwelld, P., and Bryana, A. 2008. Collaborative development of clinical templates as a national resource. *Internation Journal of Medical Informatics* 78(1):95–100.
83. Quantin, C., Binquet, C., Bourquard, K., Pattisina, R., Gouyon-Cornet, B., Ferdynus, C., Gouyon, J. B., and Allaert, F. A. 2004. A peculiar aspect of patients' safety: the discriminating power of identifiers for record linkage. *Studies in Health Technology and Informatics* 103:400–406.
84. Arellano, M. G., and Weber, G. I. 1998. Issues in identification and linkage of patient records across an integrated delivery system. *Journal of Healthcare Information Management* 12(3):43–52.
85. Cruz-Correia, R., Vieira-Marques, P., Ferreira, A., Oliveira-Palhares, E., Costa, P., and Costa-Pereira, A. 2006. Monitoring the integration of hospital information systems: how it may ensure and improve the quality of data. *Studies in Health Technology and Informatics* 121:176–82.
86. Dunn, H. L. 1946. Record linkage. *American Journal of Public Health* 36(12):1412.

87. Blakely, T., and Salmond, C. 2002. Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 31(6):1246–1252.
88. Scheuren, F. 1997. Linking health records: human rights concerns. In *Record Linkage Techniques—1997: Proceedings of an International Workshop and Exposition*, March 20–21, 1997, Arlington, VA, p. 404. Federal Committee on Statistical Methodology, Office of Management and Budget.
89. Evans, J. M. M., and MacDonald, T. M. 1999. Record-linkage for pharmacovigilance in Scotland. *British Journal of Clinical Pharmacology* 47(1):105–110.
90. Karmel, R., and Gibson, D. 2007. Event-based record linkage in health and aged care services data: a methodological innovation. *BMC Health Services Research* 7(1):154.
91. Hägglund, M., Chen, R., Scandurra, I., and Koch, S. 2009. Modeling shared care plans using CONTsys and openEHR to support shared homecare of elderly. Submitted.
92. Chen, R., Hemming, G., and Åhlfeldt, H. 2009. Representing a chemotherapy guideline using openEHR and rules. Submitted.
93. Chen, R., Klein, G., Sundvall, E., Karlsson, D., and Åhlfeldt, H. 2009. Archetype-based import and export of EHR content models: pilot experience with a regional EHR system. Submitted.
94. Pearson, R. K. 2005. *Mining Imperfect Data: Dealing with Contamination and Incomplete Records*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
95. Hawkins, D. M. 1980. *Identification of Outliers*. London: Chapman and Hall.
96. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone. C. J. 1984. *Classification and Regression Trees*. New York, NY: Chapman and Hall/CRC.
97. Rodrigues, P. P., Gama, J., and Bosnizć, Z. 2008. Online reliability estimates for individual predictions in data streams. In *Proceedings of the 8th International Conference on Data Mining Workshops (ICDMWorkshops'08)*, pp. 36–45. Pisa, Italy, December, IEEE Computer Society Press.
98. Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7:1–26.
99. Hastie, T., Tibshirani, R., and Friedman, J. 2000. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY: Springer Verlag.
100. Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–140.
101. Drucker, H. 1997. Improving regressors using boosting techniques. In *Machine Learning: Proceedings of the 14th International Conference*, pp. 107–115.
102. Gama, J., and Rodrigues, P. P. 2007. Data stream processing. In *Learning from Data Streams—Processing Techniques in Sensor Networks*, J. Gama and M. Gaber, eds., chapter 3, pp. 25–39. Berlin: Springer Verlag.
103. Gama, J., and Gaber, M., eds. 2007. *Learning from Data Streams—Processing Techniques in Sensor Networks*. Berlin: Springer Verlag.
104. Gama, J., Medas, P., Castillo, G., and Rodrigues, P. P. 2004. Learning with drift detection. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence (SBIA 2004)*, volume 3171 of *Lecture Notes in Artificial Intelligence*, A. L. C. Bazzan and S. Labidi, eds., pp. 286–295, São Luiz, Maranhão, Brazil, October 2004. Springer Verlag.
105. Böttcher, M., Höppner, F., and Spiliopoulou, M. 2008. On exploiting the power of time in data mining. *SIGKDD Explorations* 10(2):3–11.