Chapter 15

Mining the Electronic Health Record for Disease Knowledge

Elizabeth S. Chen and Indra Neil Sarkar

Abstract

The growing amount and availability of electronic health record (EHR) data present enhanced opportunities for discovering new knowledge about diseases. In the past decade, there has been an increasing number of data and text mining studies focused on the identification of disease associations (e.g., disease–disease, disease–drug, and disease–gene) in structured and unstructured EHR data. This chapter presents a knowledge discovery framework for mining the EHR for disease knowledge and describes each step for data selection, preprocessing, transformation, data mining, and interpretation/validation. Topics including natural language processing, standards, and data privacy and security are also discussed in the context of this framework.

Key words Electronic health record, Knowledge discovery in databases, Data mining, Text mining, Natural language processing, Data warehouse, Data privacy and security, Standards

1 Introduction

The increased adoption of electronic health record (EHR) systems has the potential for enhanced collection and access to a wide range of information about an individual's lifetime health status and health care to support a range of "secondary uses" including genomic, clinical, and public health research [1–6]. The use of knowledge discovery and data mining approaches for transforming health data into actionable knowledge has been an active area of research and will continue to advance in the era of "big data" [7–10]. These approaches have been applied to EHR data for studying patterns and relationships between diseases (comorbidity analysis) [11–13], drugs and adverse events (pharmacovigilance) [14–16], as well as drugs and genes (pharmacogenomics) [17]. Potential uses of this knowledge range from hypothesis generation to clinical decision support.

The knowledge discovery in databases (KDD) process is defined as "the nontrivial process of identifying valid, novel,

Vinod D. Kumar and Hannah Jane Tipney (eds.), *Biomedical Literature Mining*, Methods in Molecular Biology, vol. 1159, DOI 10.1007/978-1-4939-0709-0_15,

potentially useful, and ultimately understandable patterns in data" [18, 19]. The major steps in this interactive and iterative process are data selection, preprocessing, transformation, data mining, and interpretation/validation. While *data mining* is traditionally applied to collections of "structured" data from databases, *text mining* or *text data mining* is the application of data mining techniques to collections of "unstructured" or "semi-structured" textual data [20]. The text mining process typically involves the use of natural language processing (NLP) techniques to extract structured data from unstructured narrative for subsequent use [21].

This chapter presents a general framework for mining the EHR for disease knowledge that is adapted from the KDD and text mining processes. The sets of techniques associated with each step in this "disease knowledge discovery" (DKD) process are described along with a discussion of key issues ranging from data privacy and security to standardization. Given the breadth of methods and applications, this protocol is focused on describing an approach for identifying pairwise disease associations (e.g., disease–disease, disease–drug, and disease–gene) where the driving example is a study focused on identifying comorbidities for type 2 diabetes mellitus.

2 Materials

2.1 Electronic Health Record

An EHR system includes a "longitudinal collection of electronic health information for and about persons, where health information is defined as information pertaining to the health of an individual or health care provided to an individual" [22] (*see* **Note 1**). Typically, an EHR system interfaces with and integrates data from ancillary systems such as registration, billing, laboratory, radiology, and pharmacy [23]. Data within the EHR fall into two major categories: *structured*—discrete data elements (e.g., demographics, billing diagnoses, problems, procedures, medications, and allergies) that may be associated with codes and are "computer understandable" and *unstructured* (or *semi-structured*)—free-text narrative (e.g., clinical notes and reports) that may include some structure and can be converted into a structured form using methods such as NLP (*see* **Note 2**) [24]. The underlying real-time database for EHR systems is often referred to as a Clinical Data Repository (CDR) [25].

2.2 Data Warehouse

For reporting and data analysis purposes, a separate database or data warehouse is typically available that is populated with data through an extract, transform, and load (ETL) process [25]. Depending on the intended use (e.g., for quality or research) and contents, this is referred to as a Clinical Data Warehouse, Research Data Warehouse, Enterprise Data Warehouse, Integrated Data Warehouse, or Integrated Data Repository (*see* **Note 3**) [26]; for

the remainder of this chapter, the term Enterprise Data Warehouse (EDW) will be used. In addition to the EHR and ancillary systems, the EDW may also incorporate data from other systems such as disease registries, biobanks, payer claims databases, and literature databases. The standard process for accessing and requesting data from an EDW for research purposes involves Institutional Review Board (IRB) approval and training in the protection of human subjects in research (*see* **Note 4**).

2.3 Other Resources

In addition to the aforementioned institutional resources, there are several publicly accessible repositories of de-identified EHR data that are each associated with an approval process (e.g., requiring IRB approval or a Data Use Agreement). These include the following:

- Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) research database [27, 28]—intensive care unit data from Beth Israel Deaconess Medical Center.
- i2b2 NLP Research Datasets [29]—discharge summaries from Partners HealthCare that have been used in a series of NLP challenges (e.g., for de-identification [30], smoking status [31], and medications [32]).
- Integrating data for analysis, anonymization, and sharing (iDASH) Data Repository [33, 34]—open, community-serving, crowd-sourced resource for high-quality collections of medical data, including images and text.

3 Methods

The DKD process aligns with the DIKW hierarchy that represents the relationship between *d*ata, *i*nformation, *k*nowledge, and *w*isdom [35]. As depicted in Fig. 1, the DKD process involves the following: *Data selection*—identifying and extracting data from the EHR (through the EDW) and potentially other sources for a particular application or study; *preprocessing*—de-identifying, cleaning, and enriching the dataset, which may include the use of NLP techniques for unstructured data and applying standards for integrating structured data; *transformation*—reducing and converting the dataset in preparation for the data mining step; *data mining*—choosing and implementing the algorithms for generating patterns; and *interpretation/evaluation*—visualizing and validating the patterns, which may lead to revising and reiterating through the previous steps. The result of this process is the validation of known knowledge and perhaps more significantly the potential discovery of new disease knowledge. Key issues underlying these steps include *data privacy and security* (*see* **Note 4**) and *standards* (*see* **Note 5**).

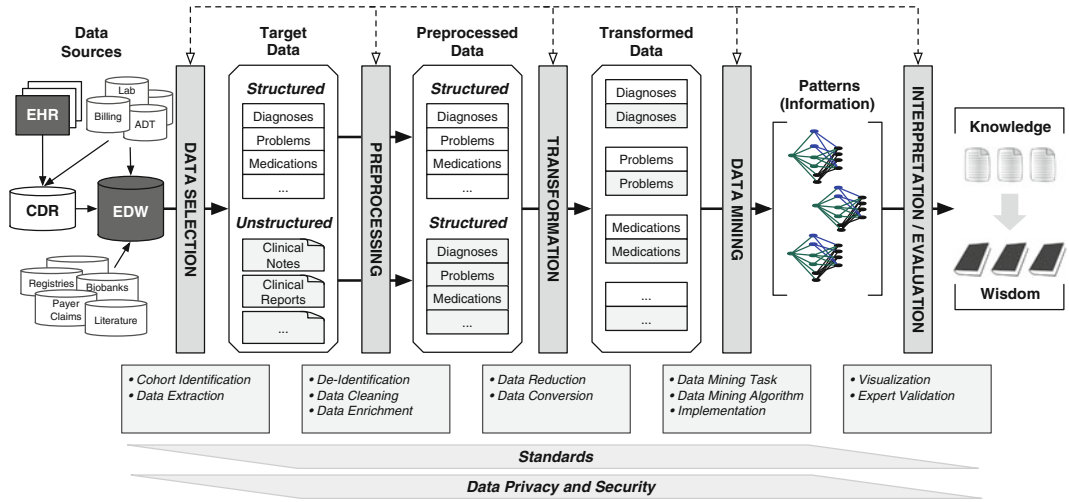


Fig. 1 Overview of disease knowledge discovery process

3.1 Example Studies

1. The driving example in this section will be a study focused on identifying comorbidities associated with type 2 diabetes mellitus (henceforth referred to as the "diabetes comorbidity study").
2. For additional examples, there are several studies involving the use of NLP to extract information from clinical notes (e.g., discharge summaries) and biomedical literature (e.g., PubMed/MEDLINE) to acquire disease–disease [11, 12], disease–finding [36, 37], disease–drug [38, 39], and disease–symptom [40, 41] relationships.
3. Other example studies have focused on structured data such as billing diagnoses, problems, procedures, medications, laboratory results, and vital signs to generate disease co-occurrences (e.g., disease–disease [13], disease–drug [42, 43], disease–lab [42, 43], disease–procedure [44], and disease–gene [45]).
4. While the aforementioned examples are focused on pairwise associations that do not take into consideration time or the sequence of events, there have been studies focused on the discovery of larger associations (e.g., triplet combinations such as disease–disease–disease [46]) and temporal associations [47], which are out of scope for this protocol.

3.2 Defining the Study

Prior to starting the DKD process, there should be a good understanding of the application of interest in order to guide each of the steps. For example, high-level questions include the following:

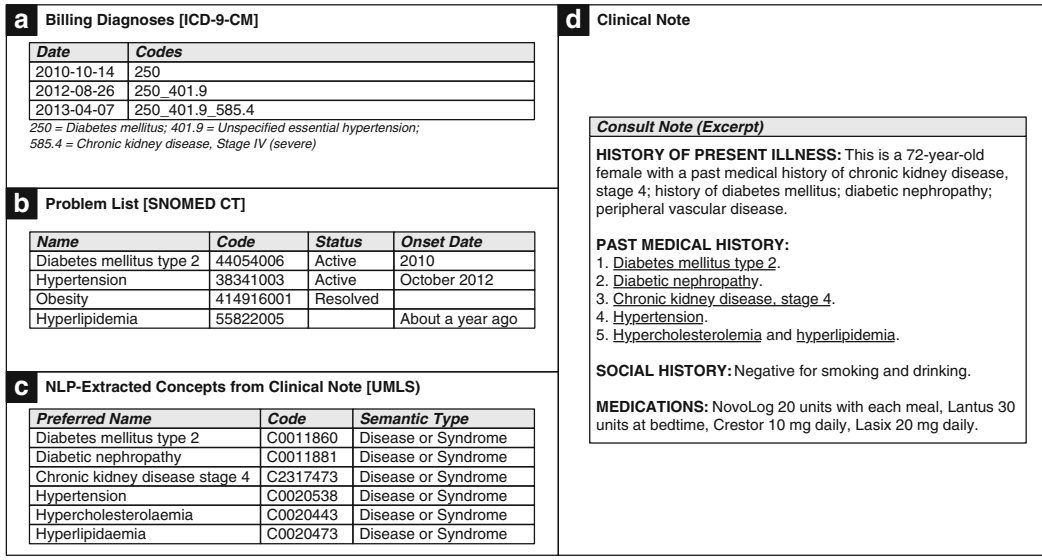
- How to access, extract, and protect the EHR data?
- What type(s) of EHR data to use (e.g., structured diagnoses or unstructured discharge summaries)?
- Are NLP techniques needed to process unstructured data?

- What standards have been used or what standards can be applied to facilitate data integration?
- How to select and identify the patient cohort (e.g., focus on particular disease(s) or specific time frame)?
- Which data mining techniques to select and implement?
- How to evaluate (e.g., involve clinical experts or compare results to medical knowledge resources)?
- What are anticipated uses of the discovered knowledge?

3.3 Data Selection

The goal of data selection is to create a target dataset by first identifying a cohort based on a set of inclusion/exclusion criteria and then extracting the requisite data. Depending on the institution, there may be resources and tools available to facilitate cohort identification and data extraction (*see* **Note 3**).

1. Criteria that can be used for *cohort identification* include patient demographics (e.g., age, sex, race/ethnicity, and zip code), encounters, diagnoses, procedures, medications, and laboratory results. These latter four types of data may be associated standardized codes such as ICD, CPT, RxNorm, and Logical Observation Identifiers Names and Codes (LOINC), respectively (*see* **Note 5**). In addition, keyword searching, regular expression matching, or NLP could enable the use of clinical notes and reports for meeting specified criteria (*see* **Note 2**).
2. The cohort identification process is often challenging due to heterogeneous data sources, lack of standards, and wealth of unstructured text. Recent efforts in "EHR-based phenotyping" have been focused on the development and validation of standardized phenotyping algorithms [48, 49]. For example, Phenotype KnowledgeBase (PheKB) [50], developed as part of the Electronic Medical Records and Genomics (eMERGE) Network [51], includes algorithms for conditions such as cataracts, dementia, and type 2 diabetes mellitus. This latter algorithm defines a set of data elements (e.g., lists of ICD-9-CM, RxNorm, and LOINC codes associated with both type 1 and type 2 diabetes mellitus), definitions (e.g., abnormal lab values such as a hemoglobin A1c $\geq$6.0 %), and logic for identifying cases and controls. Depending on the study, further criteria could be applied such as limiting to patients with encounters during a specified time period (e.g., January 1, 2012, to December 31, 2012) or a specific age group (e.g., children [<18] or adults [$\geq$18]).
3. Once the cohort has been identified, the next step is *data extraction*. This process involves determining what data are needed for those patients, what is available in the EDW, and how to obtain and securely store the resulting dataset.



a Billing Diagnoses [ICD-9-CM]

Date	*Codes*
2010-10-14	250
2012-08-26	250_401.9
2013-04-07	250_401.9_585.4

250 = Diabetes mellitus; 401.9 = Unspecified essential hypertension; 585.4 = Chronic kidney disease, Stage IV (severe)

b Problem List [SNOMED CT]

Name	*Code*	*Status*	*Onset Date*
Diabetes mellitus type 2	44054006	Active	2010
Hypertension	38341003	Active	October 2012
Obesity	414916001	Resolved	
Hyperlipidemia	55822005		About a year ago

c NLP-Extracted Concepts from Clinical Note [UMLS]

Preferred Name	*Code*	*Semantic Type*
Diabetes mellitus type 2	C0011860	Disease or Syndrome
Diabetic nephropathy	C0011881	Disease or Syndrome
Chronic kidney disease stage 4	C2317473	Disease or Syndrome
Hypertension	C0020538	Disease or Syndrome
Hypercholesterolaemia	C0020443	Disease or Syndrome
Hyperlipidaemia	C0020473	Disease or Syndrome

d Clinical Note

Consult Note (Excerpt)

HISTORY OF PRESENT ILLNESS: This is a 72-year-old female with a past medical history of chronic kidney disease, stage 4; history of diabetes mellitus; diabetic nephropathy; peripheral vascular disease.

PAST MEDICAL HISTORY:
1. Diabetes mellitus type 2.
2. Diabetic nephropathy.
3. Chronic kidney disease, stage 4.
4. Hypertension.
5. Hypercholesterolemia and hyperlipidemia.

SOCIAL HISTORY: Negative for smoking and drinking.

MEDICATIONS: NovoLog 20 units with each meal, Lantus 30 units at bedtime, Crestor 10 mg daily, Lasix 20 mg daily.

Fig. 2 Example EHR data—billing diagnoses (**a**), problem list (**b**), and NLP-extracted concepts (**c**) from clinical note (**d**; content adapted from [52])

For the diabetes comorbidity study, potential data types include *billing diagnoses* where a single patient encounter may be associated with multiple ICD codes (Fig. 2a); *problem lists* where an entry may be associated with a name and code (from ICD or other coding systems such as Systematized Nomenclature of Medicine—Clinical Terms, SNOMED CT), status (e.g., active, inactive, or resolved), onset date, and other information (Fig. 2b); and *clinical notes* that typically include information such as note type (e.g., progress note, consult note, or discharge summary) and author type (e.g., attending or resident) as part of the metadata (Fig. 2d; adapted from [52]). Each of these data types offers a unique perspective (e.g., administrative vs. clinical and structured vs. unstructured) and could be analyzed individually for comparison or collectively.

4. As part of the formal EDW data request, the request type (e.g., aggregate, de-identified, or identified dataset), data types, specific data elements, and any additional restrictions should be provided for creating the data extract (*see* **Notes 3** and **4**). These additional restrictions could include retrieving billing diagnoses for a specific time period (e.g., last 5 years), limiting to active problem list entries, and obtaining only the most recent consult note. Depending on the request type, the resulting dataset should be stored in the appropriate environment for further analysis (*see* **Note 4**).

3.4 Preprocessing

Once the dataset is obtained, major preprocessing tasks include de-identification, data cleaning, and data enrichment.

1. In some cases, the dataset may already be de-identified; however, in other cases, there may be a need to perform *de-identification* if the dataset includes any of the 18 Protected Health Information (PHI) elements as defined by the Health Insurance Portability and Accountability Act (HIPAA) in the United States (*see* **Note 4**).
2. *Data cleaning* involves a set of tasks for handling noisy, incomplete, and inconsistent data in order to improve data quality (e.g., completeness, correctness, concordance, plausibility, and currency [53]) [54]. While EHR systems include error checking and other functionalities for improving data entry in some areas, there is still the possibility for erroneous and missing values. For example, in a dataset containing ages, there may be invalid values (e.g., "1BC" or "1.5.67"), outliers (e.g., "−100.0" or "1000000") or missing or null values. Depending on the intended use, options include keeping these cases (e.g., if the study goal is to identify comorbidities for the cohort as a whole) or removing these cases (e.g., if the goal is to compare comorbidities based on particular age groups).
3. In cases where data originate from different sources, there may be variation in content and format where standards could play a key role in facilitating data integration (*see* **Note 5**). For example, use of different timestamp formats (e.g., January 1, 2013, vs. 01/01/2013) where a standard like ISO 8601 for the representation of dates and time could be used (e.g., YYYY-MM-DD or YYYYMMDD) [55]. Another example of reformatting is separating data that may be concatenated within a single field such as in Fig. 2a where each diagnosis (represented by an ICD-9-CM code) is delimited by the "_" character.
4. *Data enrichment* is focused on enhancing the dataset, which may involve NLP, standards, and other external sources such as the Unified Medical Language System (UMLS) [56]. For datasets including unstructured data, NLP techniques can be used to extract, structure, and encode disease and other information from clinical notes (either the entire note or particular sections within the note) with UMLS concepts (*see* **Notes 2** and **5**). Figure 2c depicts the UMLS concepts (represented by Concept Unique Identifiers [CUI]) and semantic types identified from the Past Medical History section of the clinical note in Fig. 2d.
5. As described earlier, billing diagnoses, problem lists, and clinical notes can each provide disease information; however, there may be challenges integrating these sources due to use of

different coding systems (e.g., local terminology vs. ICD-9-CM vs. SNOMED CT) and different levels of granularity (e.g., "Diabetes mellitus" vs. "Diabetes mellitus type 2"). Resources such as the UMLS, which contains over 150 source vocabularies and provides linkages between them (e.g., mappings between ICD-9-CM and SNOMED CT), can be leveraged to facilitate integration and potentially handle granularity issues through the use of hierarchical relationships. For example, for hypertension, ICD-9-CM code 401.9 and SNOMED CT code 38341003 are linked through UMLS CUI C0020443 (Fig. 2a–c).

3.5 Transformation

After completing the preprocessing tasks, the next step is to reduce and convert the dataset in preparation for the data mining step.

1. Among the *data reduction* strategies are aggregation, generalization, and dimensionality reduction. In the context of the diabetes comorbidity study, a particular disease may be mentioned multiple times in the billing diagnoses across encounters, represented at different levels of granularity in the problem list, and mentioned throughout different clinical notes. If the interest is only if the disease is present rather than when the disease was present (e.g., patient had disease x at time y), then these multiple occurrences could be aggregated into a single occurrence (considering time and temporality is the focus of temporal data mining [57], which is out of scope for this chapter).
2. Generalization techniques can also be applied that involve transforming specific values to more general ones using domain concept hierarchies [58]. For diseases, this could involve using customized disease classes defined by clinical experts such as the PheWAS code groups for ICD-9-CM codes [45, 59], existing classification schemes such as the Clinical Classifications Software (CCS) for grouping ICD-9-CM codes [60], or leveraging the hierarchies in terminological systems such as ICD, SNOMED CT, and UMLS (*see* **Note 5**). For example, the ICD-9-CM hierarchy can be used to generalize specific codes to 250.0 (as depicted in Fig. 3) whereas CCS categorizes 250.00 and 250.01 (along with 10 other ICD-9-CM codes) into "Diabetes mellitus without complications" and 250.02 and 250.03 (along with 55 other ICD-9-CM codes) into "Diabetes mellitus with complications."
3. Dimensionality reduction can be used to limit the number of dimensions (or features or attributes) by only selecting or transforming those that are essential to the analysis [61]. For example, this could involve excluding attributes such as the status associated with a problem or combining the status value with the problem name (e.g., "Hypertension-*Active*" or "Hypertension-*Resolved*").

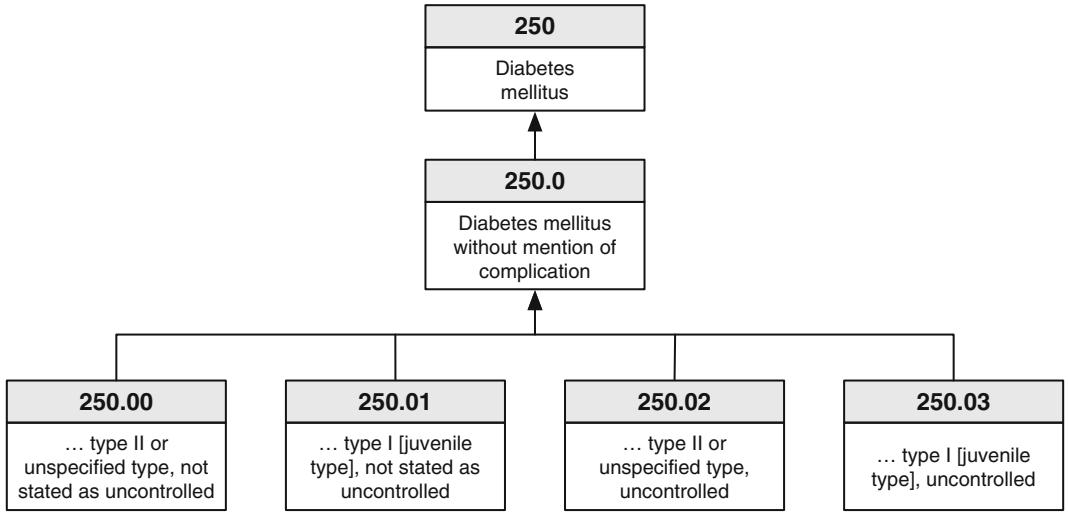


Fig. 3 Portion of ICD-9-CM hierarchy for diabetes mellitus

4. Depending on the data mining implementation, there may be specific formatting requirements for the dataset. To accommodate these and other implementation-specific requirements, *data conversion* needs to be performed. For example, the dataset may include one row per disease per patient whereas a specific data mining tool requires a single row per patient that includes the list of diseases separated by a particular delimiter (e.g., comma or tab).

3.6 Data Mining

Two primary goals of data mining are: (1) *prediction* for identifying patterns for predicting future behavior based on past behavior (e.g., as reflected in historical data) and (2) *description* for identifying patterns and relationships in data for further exploration [18, 62]. Predictive tasks include classification, regression, and time series analysis, while descriptive tasks include clustering, association rules, and sequence discovery. Given the broad array of methods and algorithms associated with each of these tasks, this section focuses on *association rules*. For comprehensive reviews of data mining methods that have been used in biomedicine and health care, *see* refs. 7, 9, 63.

1. Several open-source tools are available for performing data mining tasks and visualizing the results, including Weka and R that can be used for association rule mining [64, 65].
2. Association rule mining (or association rule learning or association rule generation) is a commonly used approach to discover interesting relationships between items in large datasets [62, 66]. An important consideration is that association rules convey co-occurrence relationships rather than causality but may

serve as a first step for generating hypotheses relating to causal relationships. Typically, association rule mining is performed on a dataset that includes a set of "transactions" (e.g., patients) where each transaction contains a subset of "items" (e.g., diseases) referred to as an "itemset." A number of algorithms exist (e.g., Apriori, Eclat, and FP-Growth [67, 68]) where the general algorithm is to first generate frequent itemsets (e.g., combinations of diseases that satisfy a specified threshold) and then generate rules in the form of $X \Rightarrow Y$ where X and Y are sets of items (e.g., {Diabetes mellitus type 2} ⇒ {Hypertension} or {Diabetes mellitus type 2, Hypertension} ⇒ {Obesity}).

3. While support and confidence are common measures for conveying the strength of each rule, there are a number of other established "interestingness" measures that can be used [69, 70]. For example, in a comparison of five common measures (support, confidence, chi-square [χ^2], interest [or lift], and conviction), χ^2 appeared to produce more accurate relationships [42] and has been the primary statistic used in several studies related to discovering disease associations in the EHR [36, 38, 41]. Other studies have used measures such as odds ratio [12, 13], relative risk [71], and relative reporting ratio [72].
4. A known challenge of association rule mining is the potentially intractable search space due to the generation of large numbers of rules that may be redundant or irrelevant [68]. Numerous algorithms and techniques exist for addressing these issues that involve filtering or pruning rules to produce a more limited and valid set of rules for further review. For example, as was described in Subheading 3.5, generalization techniques can be used to generate rules at different levels of granularity using concept hierarchies.
5. Another important consideration is ensuring that potentially important rules are not missed due to setting thresholds that are too high (e.g., for support and confidence values). This is known as the "rare item problem" where a variety of techniques have been proposed such as using multiple minimum support values [73].

3.7 Interpretation/ Evaluation

1. Different *visualization techniques* can be used to facilitate exploration of the resulting patterns. For example, the ability to review and interact with patterns represented in a graphical network can provide the "big picture" as well as allow users to focus on particular parts of the network. There are a number of open-source visualization tools (e.g., GraphViz [74] and Cytoscape [75]) as well as data mining tools that include visualization (e.g., Weka [76] or Orange [77]).
2. For assessing the validity of patterns, one or more *clinical experts* are often involved who manually review the patterns

(all or a subset) to determine if each one is "known" or "unknown" (and thus potentially novel) as well as provide any interpretations or explanations. For example, in a study focused on identifying disease-finding associations in discharge summaries, the evaluation involved having two experts categorize each association as a "direct association," "indirect association," or "non-association" [36, 37].

3. Other validation approaches involve the use of established *medical knowledge sources* such as biomedical literature (e.g., PubMed/MEDLINE and CINAHL) and medical references (e.g., Micromedex and UpToDate). In one published study, searches for supporting literature in PubMed/MEDLINE were conducted for both well-known and unknown (or poorly known) disease associations identified from free-text problem list entries [13]. Another pair of studies involved the use of the Lexi-Comp drug reference database, Mosby's Diagnostic and Laboratory Test Reference, Harrison's Principles of Internal Medicine, eMedicine, and UpToDate to validate disease–drug and disease–laboratory associations generated from structured EHR data [42, 43]. Other studies have involved comparing disease–disease and disease–drug associations generated from EHR data with those obtained from biomedical literature using text mining approaches [12, 38, 39].
4. For the diabetes comorbidity study, a search for "Diabetes Mellitus, Type 2"[mh] AND "Hypertension"[mh] returns almost 6,000 articles (as of September 18, 2013), thus suggesting support for this association in the literature. Manual review of a subset of these articles would then be required to fully validate an association between these two diseases.

4 Notes

1. *EHR Systems*: As described in the 2003 Institute of Medicine report on "Key Capabilities of an Electronic Health Record System," criteria for these systems include the following: (1) improve patient safety, (2) support the delivery of effective patient care, (3) facilitate management of chronic conditions, (4) improve efficiency, and (5) feasibility of implementation [22]. To address these criteria, eight categories of core functionality for an EHR system are defined: (1) health information and data, (2) result management, (3) order entry/management, (4) decision support, (5) electronic communication and connectivity, (6) patient support, (7) administrative process, and (8) reporting and population health management [22]. Options for EHR implementation range from homegrown (e.g., StarChart/StarPanel at Vanderbilt University

Medical Center [78]) to commercial (e.g., Cerner [79] and Epic [80]) to open-source (e.g., OpenVista [81]) systems. While organizations such as Health Level 7 (HL7) [82] have been actively involved with defining and developing EHR standards, there is often variation in features and functions as EHR products evolve where particular systems may be more advanced in some areas than others [25]. In addition, an institution may have multiple EHR systems (e.g., one vendor for inpatient and another for outpatient) and institution-specific customizations may lead to variations in how and where data are collected. In considering secondary uses such as research, the heterogeneous nature of EHR systems needs to be considered and accounted for.

2. *NLP*: Automated methods such as NLP enable a new level of functionality by providing a means to extract, encode, and structure relevant information from free text in a timely fashion [83]. With respect to healthcare data, the development and use of NLP tools can facilitate the capture and exchange of essential information for a variety of subsequent uses ranging from patient care to quality reporting to research. A variety of NLP algorithms and systems have emerged over the last few decades to support the tasks of extracting information captured within clinical (e.g., discharge summaries and progress notes) and biomedical (e.g., literature) text for EHR enrichment, decision support, surveillance, terminology management, and text mining [84]. Many of these systems are focused on extracting named entities (e.g., diseases, medications, or procedures), identifying contextual information such as negation and temporality (e.g., "*no* hypertension" and "myocardial infarction *five years ago*"), and may also provide mappings to an appropriate standard such as ICD-9-CM, SNOMED CT, and UMLS [85, 86]. This standardized encoding of information provides domain knowledge and is essential for promoting interoperability and facilitating subsequent analyses such as data mining (*see* **Note 5**). Example systems include Medical Language Extraction and Encoding (MedLEE) system developed at Columbia University [87–89], MetaMap from the National Library of Medicine [90], clinical Text Analysis and Knowledge Extraction System (cTAKES) released by the Open Health Natural Language Processing Consortium [91], and Health Information Text Extraction (HITEx) that is part of the i2b2 framework [92] (*see* **Note 3**).
3. *Data Warehouses*: In recent years there has been increased attention to the development and use of data warehouses that integrate administrative, clinical, biomedical, and research data from EHR and other systems. Depending on the institution, different names may be used such as Clinical Data

Warehouse, Research Data Warehouse, EDW, Integrated Data Warehouse, or Integrated Data Repository [26]. One prominent example is the i2b2 (Informatics for Integrating Biology and the Bedside) framework that has been adopted by over 60 academic health centers nationally and internationally [93]. The open-source i2b2 platform can be used for de-identified cohort discovery and hypothesis testing by researchers as well as for creation of identified datasets with the requisite approvals (*see* **Note 4**). To enable federated queries between i2b2 instances, Shared Health Research Information Network (SHRINE) implementations can be created for data sharing and cohort identification across institutions [94]. In addition to i2b2, other examples include the Synthetic Derivative at Vanderbilt University [95], Stanford Translational Research Integrated Database Environment (STRIDE) at Stanford University [96], the Enterprise Data Trust at Mayo Clinic [97], Biomedical Translational Research Information System (BTRIS) at the National Institutes of Health [98], and Translational Research Informatics and Data Management Grid (TRIAD) at Ohio State University [99]. Many of these systems provide an interface that allows researchers to first search for cohorts based on specified criteria (e.g., demographics, ICD diagnosis codes, and CPT procedure codes). After a cohort has been identified, there is typically a formal request and approval process for obtaining a data extract containing de-identified or identified data (*see* **Note 4**).

4. *Data Privacy and Security*: With the increased sharing and use of electronic health data for research purposes, attention to data privacy and security issues is essential [100–103]. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) is aimed at protecting the privacy of an individual's health information [104]. HIPAA includes two major rules: (1) Privacy Rule that governs the use and disclosure of PHI and (2) Security Rule that specifies a series of administrative, physical, and technical safeguards. Following the HIPAA Privacy Rule, a "de-identified" dataset is one where the 18 defined PHI elements (e.g., names, addresses, dates, ages, and unique identifying numbers) are removed using either the safe harbor method or expert determination while a "limited" dataset involves removal of 16 of these elements and maintains dates and geographic information [105, 106]. For datasets including unstructured data (e.g., clinical notes), text de-identification approaches are needed to remove or mask these PHI elements [107]. While a de-identified or a limited dataset typically include codes or "pseudonyms" for enabling re-linkages to individuals if needed, an "anonymized" dataset does not maintain any linkages [108]. However, despite

the use of de-identification and anonymization techniques, there may still be risks of misuse and re-identification [109]. Thus, researchers have the ethical and legal obligation as well as the responsibility to adhere to both local and federal regulations in the protection of human subjects and responsible conduct of research, including receiving the requisite training and conducting research in a HIPAA-compliant environment. For institutional data warehouses (*see* **Note 3**), there are typically several categories of requests that align with different levels of privacy: aggregate counts, limited dataset, de-identified dataset, and identified dataset [110]. Each of these request types may be associated with a set of required approvals (e.g., IRB) and forms (e.g., Data Use Agreement [DUA]).

5. *Standards*: To support semantic interoperability, there is a need for common representations and use of vocabulary standards for specifying the syntax and semantics of health information [111, 112]. Vocabulary standards include International Classification of Diseases (ICD) for diagnoses where the transition from ICD-9-CM (Ninth Revision, Clinical Modification) to ICD-10 is under way [113]; Current Procedural Terminology (CPT) for procedures [114]; LOINC for laboratory and other clinical observations [115, 116]; RxNorm for clinical drugs and drug delivery devices [117, 118]; SNOMED CT for comprehensive coding of health information [119]; and Medical Subject Headings (MeSH) for indexing biomedical literature [120]. The aforementioned standards are among the 150 source vocabularies integrated within the UMLS Metathesaurus, developed at the National Library of Medicine, that includes millions of concepts and their relationships, thus providing linkages between the source vocabularies [56, 121]. Complementary resources such as the BioPortal maintained at the National Center for Biomedical Ontology (NCBO) offer additional concepts from those vocabularies that may not be part of the UMLS [122, 123].

Acknowledgment

The example clinical note in Fig. 2d was obtained with permission from MTSamples (http://www.mtsamples.com). This work was supported in part by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Institute of Medicine (U.S.), Committee on Improving the Patient Record (eds), Dick RS, Steen EB, Detmer DE (1997) The computer-based patient record: an essential technology for health care. Revised edition. National Academy Press, Washington, DC
2. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM (2007) Bridging the inferential gap: the electronic health record and clinical evidence. Health Aff (Millwood) 26: w181–w191
3. Kohane IS (2011) Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 12:417–428
4. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C et al (2013) Electronic health records: new opportunities for clinical research. J Intern Med 274(6):547–560
5. Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, Mortoti S et al (2007) Redesigning electronic health record systems to support public health. J Biomed Inform 40:398–409
6. Denny JC (2012) Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol 8:e1002823
7. Bath P (2004) Data mining in health and medical information. Annu Rev Inform Sci Technol 38:331–369
8. van Bemmel JH, van Mulligen EM, Mons B, van Wijk M, Kors JA, van der Lei J (2006) Databases for knowledge discovery. Examples from biomedicine and health care. Int J Med Inform 75:257–267
9. Iavindrasana J, Cohen G, Depeursinge A, Muller H, Meyer R, Geissbuhler A (2009) Clinical data mining: a review. Yearb Med Inform:121–133
10. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. JAMA 309:1351–1352
11. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T et al (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 7:e1002141
12. Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R (2011) Discovering disease associations by integrating electronic clinical data and medical literature. PLoS One 6:e21132
13. Hanauer DA, Rhodes DR, Chinnaiyan AM (2009) Exploring clinical associations using '-omics' based enrichment analyses. PLoS One 4:e5203
14. Wilson AM, Thabane L, Holbrook A (2004) Application of data mining techniques in pharmacovigilance. Br J Clin Pharmacol 57: 127–134
15. Wang X, Hripcsak G, Markatou M, Friedman C (2009) Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 16:328–337
16. Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C (2011) Biclustering of adverse drug events in the FDA's spontaneous reporting system. Clin Pharmacol Ther 89:243–250
17. Wilke RA, Xu H, Denny JC, Roden DM, Krauss RM, McCarty CA et al (2011) The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther 89:379–386
18. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17:37–54
19. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39:27–34
20. Hearst M (1999) Untangling text data mining. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on computational linguistics, pp 3–10
21. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB (2007) Frontiers of biomedical text mining: current progress. Brief Bioinform 8:358–375
22. Institute of Medicine (2003) Key capabilities of an electronic health record system. National Academies Press, Washington, DC
23. National Institutes of Health National Center for Research Resources and MITRE Corporation (2006) Electronic health records overview. http://www.himss.org/files/HIMSSorg/content/files/Code%20180%20MITRE%20Key%20Components%20of%20an%20EHR.pdf
24. ASTM Standard E1384 (2013) Standard guide for content and structure of the Electronic Health Record (EHR). ASTM International, West Conshohocken, PA
25. Carter J (2008) Electronic health records for clinicians and administrators: infrastructure and supporting technologies. In: Carter J (ed) Electronic health records, 2nd edn. American College of Physicians, Philadelphia, PA
26. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N (2012) Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc 19: e119–e124
27. http://mimic.physionet.org/
28. Scott DJ, Lee J, Silva I, Park S, Moody GB, Celi LA et al (2013) Accessing the public

MIMIC-II intensive care relational database for clinical research. BMC Med Inform Dec Mak 13:9
29. https://i2b2.org/NLP/DataSets/
30. Uzuner O, Luo Y, Szolovits P (2007) Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 14:550–563
31. Uzuner O, Goldstein I, Luo Y, Kohane I (2008) Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 15:14–24
32. Uzuner O, Solti I, Cadag E (2010) Extracting medication information from clinical text. J Am Med Inform Assoc 17:514–518
33. Ohno-Machado L, Bafna V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K et al (2012) iDASH: integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc 19:196–201
34. http://idash.ucsd.edu/data-repository-0
35. Ackoff R (1989) From data to wisdom. J Appl Syst Anal 16:3–9
36. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G (2005) Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc:106–110
37. Cao H, Hripcsak G, Markatou M (2007) A statistical methodology for analyzing co-occurrence data from a large sample. J Biomed Inform 40:343–352
38. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C (2008) Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 15:87–98
39. Chen ES, Stetson PD, Lussier YA, Markatou M, Hripcsak G, Friedman C (2007) Detection of practice pattern trends through Natural Language Processing of clinical narratives and biomedical literature. AMIA Annu Symp Proc:120–124
40. Wang X, Hripcsak G, Friedman C (2009) Characterizing environmental and phenotypic associations using information theory and electronic health records. BMC Bioinforma 10(Suppl 9):S13
41. Wang X, Chase H, Markatou M, Hripcsak G, Friedman C (2010) Selecting information in electronic health records for knowledge acquisition. J Biomed Inform 43:595–601
42. Wright A, Chen ES, Maloney FL (2010) An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 43: 891–901
43. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ et al (2011) A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc 18:859–867
44. Doddi S, Marathe A, Ravi SS, Torney DC (2001) Discovery of association rules in medical data. Med Inform Internet Med 26: 25–33
45. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K et al (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26:1205–1210
46. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG et al (2006) Data mining and clinical data repositories: insights from a 667,000 patient data set. Comput Biol Med 36:1351–1377
47. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R (2011) Mining health care administrative data with temporal association rules on hybrid events. Methods Inf Med 50: 166–179
48. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V et al (2013) Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 20:e147–e154
49. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG et al (2013) Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 20(e2):e226–e231
50. http://www.phekb.org/
51. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA et al (2013) The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 15(10):761–771
52. http://www.mtsamples.com/site/pages/sample.asp?type=97-Consult%20-%20History%20and%20Phy.&sample=2063-Gen%20Med%20Consult%20-%2049
53. Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 20: 144–151
54. Rahm E, Do H (2000) Data cleaning: problems and current approaches. IEEE Data Eng Bull 23:3–13
55. http://www.w3.org/TR/NOTE-datetime
56. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:D267–D270
57. Post AR, Harrison JH Jr (2008) Temporal data mining. Clin Lab Med 28:83–100, vii

58. Carter C, Hamilton H (1995) A fast, on-line generalization algorithm for knowledge discovery. Appl Math Lett 8:5–11
59. http://knowledgemap.mc.vanderbilt.edu/research/content/phewas
60. http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
61. Liu H, Motoda H (1998) Feature extraction, construction and selection: a data mining perspective. Kluwer, Boston
62. Dunham MH (2003) Data mining introductory and advanced topics. Prentice Hall, Upper Saddle River, NJ
63. Sarkar IN (2013) Methods in biomedical informatics: a pragmatic approach, 1st edn. Academic, New York
64. Zupan B, Demsar J (2008) Open-source tools for data mining. Clin Lab Med 28:37–54, vi
65. http://www.kdnuggets.com/software/index.html
66. Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining, 1st edn. Pearson Addison Wesley, Boston
67. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. Proceedings of the 20th International conference on very large data bases, pp 487–499
68. Hipp J, Guntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining—a general survey and comparison. ACM SIGKDD Explor Newslett 2:58–64
69. Tan P, Kumar V, Srivastava J (2002) Selecting the right interestingness measure for association patterns. Proceedings of the 8th ACM SIGKDD International conference on knowledge discovery and data mining, pp 32–41
70. Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T (2007) Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artif Intell Med 41:177–196
71. Hidalgo CA, Blumm N, Barabasi AL, Christakis NA (2009) A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 5:e1000353
72. Harpaz R, Chase HS, Friedman C (2010) Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinforma 11(Suppl 9):S7
73. Liu B, Hsu W, Ma Y (1999) Mining association rules with multiple minimum supports. KDD '99 Proceedings of the 5th ACM SIGKDD International conference on knowledge discovery and data mining, pp 337–341
74. http://www.graphviz.org/
75. http://www.cytoscape.org/
76. http://www.cs.waikato.ac.nz/ml/weka/
77. http://orange.biolab.si/
78. http://informatics.mc.vanderbilt.edu/archives/starchart
79. http://cerner.com/
80. http://www.epic.com/
81. http://medsphere.com/vista-to-openvista
82. http://www.hl7.org
83. Friedman C, Johnson S (2006) Natural language and text processing in biomedicine. In: Shortliffe E, Cimino JJ (eds) Biomedical informatics computer applications in health care and biomedicine, 3rd edn. Springer, New York
84. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF (2008) Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform:128–144
85. Cimino JJ (1996) Review paper: coding systems in health care. Methods Inf Med 35:273–284
86. Cimino JJ, Zhu X (2006) The practical impact of ontologies on biomedical informatics. Yearb Med Inform:124–135
87. Friedman C (2000) A broad-coverage natural language processing system. Proc AMIA Symp:270–274
88. Friedman C, Hripcsak G, Shagina L, Liu H (1999) Representing information in patient reports using natural language processing and the extensible markup language. J Am Med Inform Assoc 6:76–87
89. Friedman C, Shagina L, Lussier Y, Hripcsak G (2004) Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 11:392–402
90. Aronson AR, Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 17:229–236
91. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC et al (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17:507–513
92. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R (2006) Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Dec Mak 6:30
93. Kohane IS, Churchill SE, Murphy SN (2012) A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc 19:181–185
94. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J et al (2013) SHRINE: enabling nationally scalable multisite disease studies. PLoS One 8:e55811

95. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR et al (2008) Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 84:362–369
96. Lowe HJ, Ferris TA, Hernandez PM, Weber SC (2009) STRIDE—an integrated standards-based translational research informatics platform. AMIA Annu Symp Proc 2009:391–395
97. Chute CG, Beck SA, Fisk TB, Mohr DN (2010) The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 17:131–135
98. Cimino JJ, Ayres EJ (2010) The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform 160:1299–1303
99. Payne P, Ervin D, Dhaval R, Borlawsky T, Lai A (2011) TRIAD: the Translational Research Informatics and Data Management Grid. Appl Clin Inform 2:331–344
100. Wylie JE, Mineau GP (2003) Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 21:113–116
101. Malin B, Karp D, Scheuermann RH (2010) Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med 58: 11–18
102. Krishna R, Kelleher K, Stahlberg E (2007) Patient confidentiality in the research use of clinical medical databases. Am J Public Health 97:654–658
103. Berman JJ (2002) Confidentiality issues for medical data miners. Artif Intell Med 26: 25–36
104. http://www.hhs.gov/ocr/privacy/index.html
105. Gunn PP, Fremont AM, Bottrell M, Shugarman LR, Galegher J, Bikson T (2004) The Health Insurance Portability and Accountability Act Privacy Rule: a practical guide for researchers. Med Care 42:321–327
106. Nosowsky R, Giordano TJ (2006) The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med 57:575–590
107. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH (2010) Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 10:70
108. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K (2012) Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care 50(Suppl): S82–S101
109. El Emam K, Jonker E, Arbuckle L, Malin B (2011) A systematic review of re-identification attacks on health data. PLoS One 6:e28071
110. Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I (2011) Strategies for maintaining patient privacy in i2b2. J Am Med Inform Assoc 18(Suppl 1):i103–i108
111. Hammond WE (2005) The making and adoption of health data standards. Health Aff (Millwood) 24:1205–1213
112. Chen ES, Melton GB, Sarkar IN (2012) Translating standards into practice: experiences and lessons learned in biomedicine and health care. J Biomed Inform 45:609–612
113. http://www.who.int/classifications/icd/en/
114. http://www.ama-assn.org/go/cpt
115. http://loinc.org/
116. Vreeman DJ, McDonald CJ, Huff SM (2010) LOINC(R)—a universal catalog of individual clinical observations and uniform representation of enumerated collections. Int J Funct Inform Personal Med 3:273–291
117. http://www.nlm.nih.gov/research/umls/rxnorm/
118. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R (2011) Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 18:441–448
119. http://www.ihtsdo.org/snomed-ct/
120. http://www.nlm.nih.gov/mesh/
121. http://www.nlm.nih.gov/research/umls/
122. http://bioportal.bioontology.org/
123. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N et al (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 37: W170–W173