Clinical data reuse or secondary use: current status and potential future progress

Meystre, S M; Lovis, Christian; Bürkle, T; Tognola, G; Budrionis, A; Lehmann, C U

How to cite

MEYSTRE, S M et al. Clinical data reuse or secondary use: current status and potential future progress. In: Yearbook of Medical Informatics, 2017, vol. 26, n° 1, p. 38–52. doi: 10.15265/IY-2017-007

This publication URL: https://archive-ouverte.unige.ch/unige:99671

Publication DOI: 10.15265/IY-2017-007

Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress

S. M. Meystre[a], C. Lovis[b], T. Bürkle[c], G. Tognola[d], A. Budrionis[e], C. U. Lehmann[f]

[a] Medical University of South Carolina, Charleston, SC, USA
[b] Division of Medical Information Sciences, University Hospitals of Geneva, Switzerland
[c] University of Applied Sciences, Bern, Switzerland
[d] Institute of Electronics, Computer and Telecommunication Engineering, Italian Natl. Research Council IEIIT-CNR, Milan, Italy
[e] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[f] Departments of Biomedical Informatics and Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA

Summary

Objective: To perform a review of recent research in clinical data reuse or secondary use, and envision future advances in this field.
Methods: The review is based on a large literature search in MEDLINE (through PubMed), conference proceedings, and the ACM Digital Library, focusing only on research published between 2005 and early 2016. Each selected publication was reviewed by the authors, and a structured analysis and summarization of its content was developed.
Results: The initial search produced 359 publications, reduced after a manual examination of abstracts and full publications. The following aspects of clinical data reuse are discussed: motivations and challenges, privacy and ethical concerns, data integration and interoperability, data models and terminologies, unstructured data reuse, structured data mining, clinical practice and research integration, and examples of clinical data reuse (quality measurement and learning healthcare systems).
Conclusion: Reuse of clinical data is a fast-growing field recognized as essential to realize the potentials for high quality healthcare, improved healthcare management, reduced healthcare costs, population health management, and effective clinical research.

Keywords

Medical informatics, electronic health records, health services research, health care evaluation mechanisms, clinical studies as topic

Yearb Med Inform 2017:38-52
http://dx.doi.org/10.15265/IY-2017-007
Published online May 8, 2017

1 Introduction

The growing adoption of Electronic Health Records (EHRs) in the U.S. healthcare system [1] and worldwide [2] fuels a fast growth of clinical data available in electronic format. This growth offers tremendous potential for the use of clinical data beyond its primary intent (i.e., patient care and healthcare operations). Secondary use (or reuse) of clinical data is defined as "non-direct care use of personal health information including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities."[3] Reuse of clinical data is essential to fulfill the promises for high quality healthcare, improved healthcare management, reduced healthcare costs, population health management, and effective clinical research. The existing and often biased and underspecified diagnostic and procedure codes assigned for reimbursement and administrative purposes are the easiest to reuse but are insufficient for policymakers, public health officials, funding agencies, scientists, clinicians, citizens, and industry, who need accurate and detailed clinical information, as found in patients' EHRs. Access to rich and detailed clinical information on diagnoses, treatments, and outcomes is also required for the Positive Predictive Value Medicine proposed by the U.S. National Academy of Sciences [4]. Further, the U.S. National Health Information Infrastructure (NHII) roadmap suggests that "…a comprehensive set of Patient Medical Record Information (PMRI) standards can move the Nation closer to a healthcare environment where clinically specific data can be captured once at the point of care with derivatives of this data available for meeting the needs of payers, healthcare administrators, clinical research, and public health. This environment could significantly reduce the administrative and data capture burden on clinicians; dramatically shorten the time for clinical data to be available for public health emergencies and for traditional public health purposes; profoundly reduce the cost for communicating, duplicating, and processing healthcare information; and, last but not least, greatly improve the quality of care and safety for all patients."[5]

Early clinical data reuse efforts often consisted of electronic databases, with manual entry of clinical data from patient paper charts. A good example was the ARAMIS databank founded in 1974, a consortium of North American rheumatic disease data banks used for multiple clinical trials [6]. This manual transcription from paper to electronic databases was time-consuming, error prone, and costly. Several EHR systems already existed at that time, but their rarity and the aforementioned costly manual translation of data strongly limited clinical data reuse efforts for many years. The development of the electronic submission of diagnostic and procedure codes, as required for Medicare and Medicaid reimbursement in the U.S. since 2003 [7], strongly enhanced the availability of this information in electronic format, and these codes quickly became the only electronic clinical data that were routinely reused.

In the past five years, the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [8] resulted in a dramatic increase in EHR implementation and use in U.S. hospitals and physician offices [9], and in large quantities of electronic clinical information becoming available in electronic format, a very appealing prospect for clinical data reuse. Incentives also spurred adoption of EHRs by general practitioners in the U.K.[10] Recent initiatives such as EHR4CR, [11] the Clinical and Translational Science Awards (CTSA) [12], the Strategic Health IT Advanced Research Projects (SHARP) program [13], and the Electronic Medical Records and Genomics (eMERGE) consortium [14] further added to this opportunity, contributing to the surge in clinical data reuse projects and publications observed. The fast growing quantity of clinical information available in electronic format makes reused clinical data a candidate for "big data" solutions [15]. As defined by Gartner, "Big data is high-volume, -velocity, and -variety information assets that demand cost-effective and innovative forms of information processing for enhanced insight and decision making" [16]. Massive quantities of unstructured data (e.g., images, scanned documents, narrative text clinical notes) from various sources and formats can be analyzed in their native state and integrated with structured data in real-time [17] to generate new information and knowledge that can then be delivered as "small data" (limited volume, in batches or near real-time, and structured) for patient-specific analysis and decision support.

The objective of this paper is to perform a review of recent research in clinical data reuse or secondary use, and envision future progress in this field.

II Materials and Methods

A Study Setting and Materials Selection

This review is based on an extensive literature search in several databases: MEDLINE (through PubMed), conference proceedings, and the ACM Digital Library. Keywords used for querying these databases included all permutations of 'reuse' or "secondary use" with "clinical data," "clinical information," "electronic health record," 'EHR,' "electronic medical record", 'EMR', "patient record," "medical record," or "clinical record." Databases were queried in February 2016. Our review focused only on research published recently (between 2005 and early 2016) in English language. We also added topic-specific publications referenced in papers that were already included.

B Selected Materials Review

Each selected publication was reviewed by the authors, and a structured analysis and summarization of its content was created and added to this review. The objective of this review was to provide readers with a large overview of published clinical data reuse research since 2005, without aiming at providing a comprehensive review of all publications in this field.

III Results

A Study Setting and Materials Selection

The initial literature search produced 359 publications (282 publications from MEDLINE and 77 distinct publications from the ACM Digital Library) using the criteria described above. After a manual examination of these publication abstracts, 35 were considered irrelevant and removed from the set, leaving 324 publications for further review. This detailed review was realized by each of the authors, focusing on specific sections and topics presented below.

B Motivations and Challenges for Clinical Data Reuse

The benefits of reusing clinical data have been well recognized for decades [3,18-21] and a detailed study by PricewaterhouseCoopers explained how reuse could enable improvements of health outcomes and costs [22]. To improve healthcare management and quality, clinical data has already been reused to measure and improve quality [23,24], predict patients length of stay, discharge, readmission, and death [25-28], and improve infection control [29-31]. Data has also been reused for early detection of diseases, pharmacovigilance, and post-market and public health surveillance [32]. In clinical research, data has been reused to accelerate and increase patient recruitment in trials [33], enable in-silico hypothesis testing [34], and enable faster and cheaper access to a richer variety of clinical information for various types of clinical research applications such as comparative effectiveness research and patient phenotype combination with genomic data. As discussed by Corevits and colleagues, clinical data reuse "will optimize research and development platforms, processes, and timelines", will generate "high-quality clinical evidence faster through better protocol feasibility assessment, improved patient identification and recruitment, and more efficient clinical study conduct, including for reporting serious adverse events", "will maximize the value to customers and diversify revenue streams" of research organizations, and enable the participation of clinical investigators and physicians in a larger number of clinical trials [35]. This topic is discussed in more detail in the Clinical Data Reuse Examples given below. Combining biomedical knowledge with reused clinical data is required for rapid "learning health systems" that would accelerate the "progression of knowledge from the laboratory bench to the patient's bedside and provide a cornerstone for health care reform."[36] This topic will also be addressed in more detail below. Clinical data reuse also offers important commercial value [37]. Clinical data is used by public and private payers for cost-effectiveness research and assistance with optimal reimbursement decisions; healthcare organizations store increasing quantities of clinical data for internal applications realizing that this data could soon become a very valuable asset. For the healthcare IT industry, research platforms allowing clinical data reuse open new business opportunities facilitated by sustainable business models [35].

Although offering multiple potential advantages, reuse of clinical data also faces multiple challenges from the observational and clinically-motivated data collection process, data quality issues, data integration and interoperability limitations, and socio-organizational constraints [21, 38-40]. Clinical data are collected for clinical use and for billing purposes. These observational data (rather than experimental data) are more process-related and frequently lack outcome data needed for effective research [21]. Clinical data are also biased by the incentives for clinicians to "upcode", by the non-random assignment of treatments, by systematic differences between patients and the general population, by the healthcare system complexity causing multiple confounders, and the large variability of measurement instruments and methods [40, 41]. The quality of data is often problematic or insufficient for research applications [42-44]. Data are often incomplete (e.g., outcomes are frequently missing) [45] or simply not randomly complete [46], patient records are fragmented, data entry errors are common, and the timeliness or currency of the data can be difficult to establish. These limitations have motivated several research teams to propose approaches for data quality assessment [47-49].

Reuse of clinical data typically implies combining heterogeneous and multidimensional sets of data into common repositories, data warehouses, or networks, with challenges in integration, interoperability, and shared meaning [21]. This topic is discussed further below. Among socio-organizational constraints, patient privacy, data ownership, intellectual property, and organizational incentives and policies are the most important. Clinical data reuse for research purposes is inevitably challenged both by legal and ethical considerations, trying to find a balance enabling scientific research within a framework in which the privacy of patients is protected [3, 50, 51]. Finally, the sale of clinical data remains an unresolved policy issue [3, 21, 52].

Recognizing the multiple potential benefits of clinical data reuse, but also the numerous aforementioned difficulties, several organizations and researchers have proposed recommendations for successful (or at least informed) clinical data reuse. The American Medical Informatics Association has published a white paper listing recommendations for a national framework for the secondary use of clinical data [3]. A similar European initiative proposed recommendations for the trustworthy reuse of health data [52], and Hersh and colleagues published recommendations [53] and caveats for clinical data reuse in comparative effectiveness research [54].

C Privacy and Ethical Concerns Related to Clinical Data Reuse

While in most countries, consent is not legally required to collect clinical patient data and in most U.S. states (except New Hampshire) patients do not legally own their medical data [55], from an ethical standpoint, patients consent indirectly to the collection, storage, transmission, access, and manipulation of their data in EHRs because they perceive the direct benefit of such data for their own care. For example, the ability of an EHR to reduce drug-drug or drug-allergy adverse events [56] or to avoid having to repeat the same medical history to every new provider [57] are tangible benefits to patients which lead to their consent for their data to be collected in the first place and then reused. While some patients express altruistic intentions and want their data to be used "so that another person might be helped," in general such behavior may not be assumed. Most advantages of data reuse benefit others (e.g., payers, providers, researchers, politicians, and society at large), than the patient. Thus, ethically, it is mandatory that the originator (from an ethical point of view which may be different than the legal point of view) and the original owner of the data - the patient - who may not be the direct beneficiary of the data reuse be properly protected in her/his rights. Table 1 explores general principles of informatics ethics applicable to clinical data reuse.

Table 1 General principles of informatics ethics (adopted from the IMIA Code of Ethics for Health Information Professionals[58]) and their impact on data reuse

Principles	Definitions	Impact on Reuse
Principle of Information-Privacy and Disposition	The fundamental right of a person to privacy and with it the right to control data about her/himself including the collection, storage, transmission, access, modification, disposition, and most importantly use of the data.	▪ Reasonable protection against any disclosure of patient data. ▪ Patient right to have data expunged or modified.
Principle of Openness	The collection, storage, transmission, access, modification, disposition, and use of a person's data must be disclosed to the person in an appropriate and timely fashion.	▪ Required notification of patients (and raising of awareness) that their data are collected and stored, transmitted, modified, and reused.
Principle of Security	Collected data must be protected by all reasonable and appropriate measures against loss, degradation, unauthorized access or destruction, use, manipulation, modification, or transmission.	▪ Security for systems allowing secondary use of data must be at or above the level of security provided for systems designed for the original use.
Principle of the Least Intrusive Alternative	Any infringement of privacy rights or the individual's right to control her/his data may only occur in the least intrusive fashion and with a minimum of interference with the rights of the affected person.	▪ Required analysis of planned reuse of data to avoid infringement or more than minimal interference.
Principle of Accountability	Any infringement of privacy rights or of the individual's right to control her/his data must be justified to the affected person in a timely manner and in an appropriate fashion.	▪ Violations of the above principles require the individuals working with reused data (not the primary data collectors) to disclose such events.

In the United States, the confidentiality of patient data is protected by the 1996 Health Insurance Portability and Accountability Act (HIPAA), the 2000 Privacy Rule (codified as 45 CFR §160 and 164) [59], and the Common Rule [60]. In the European Union, the European Convention on Human Rights and the Data Protection Directive Article 8 (95/46/ EC [61]) offer similar legal bases, with corresponding national legislations in each member states (e.g., Data Protection Act 1998 (DPA) in the UK [62]). These laws typically require the informed consent of the patient and approval of the Internal Review Board (IRB) to reuse data for research purposes. The informed consent requirement is sometimes extremely difficult or even impossible to fulfill (e.g., retrospective studies of large patient populations who moved, changed healthcare system, or died). This requirement can be waived if data is "de-identified". For clinical data to be considered de-identified, the HIPAA act and Privacy Rule require either that there is only a very small risk that the information could be used to identify the individual, subject of the information, ("Expert determination" method) or that 18 protected health information (PHI) identifiers are removed ("Safe Harbor" method) [59]. A meaningless identifier can be retained to permit re-identification of the de-identified data by a Honest Broker. The terms "anonymization" and "de-identification" are often used interchangeably, but de-identification only means that explicit identifiers are hidden or removed, while anonymization implies that the data cannot be linked to identify the patient and addresses all data, not only identifiers (i.e., de-identified data can be far from anonymous). Pseudonymization and scrubbing are two synonyms for de-identification.

The de-identification of structured data typically consists in removing or replacing data in each of the 18 PHI categories. Several commercial applications currently offer this functionality in databases (e.g., IBM Optim Data Privacy Solution, Oracle Data Masking Pack). Applications to research and public health networks [63, 64] or as a service based on the ISO 13606 EHR semantic interoperability standard [65] are examples requiring more complex implementations. Besides PHI removal or replacement, de-identification can also be achieved by segmenting [66] or 'disassociating' patient records [67]. De-identifying unstructured clinical text is a far more complex endeavor because of the difficulty to identify PHI in text [68]. It is often realized manually and requires significant resources [69]. For more scalable approaches, several authors have investigated automated text de-identification based on natural language processing (NLP) [70] using various methods. Methods are usually based on pattern matching and dictionaries, or on machine learning algorithms. Some are more generalizable than others, and certain methods perform better with some types of PHI than others [71, 72]. Recent examples such as MIST [73], BoB [74], Anonym [75], and several systems developed for the i2b2 NLP challenges [76, 77], allow for good accuracy and very limited impact on clinical information.[78] Replacing PHI with realistic surrogates [79] and adding biomedical scientific literature text [80] allowed for improved performance. Applications to French [81, 82] and Swedish [83] clinical texts have shown good or promising performance.

The anonymization of structured data has been realized with a variety of algorithms such as k-anonymity [84] or l-diversity [85] to learn useful information about a population but none about an individual, reaching ε-differential privacy [86] or other privacy protection definitions. El Emam and colleagues authored a good overview of anonymization [51]. A good detailed review of anonymization algorithms was authored by Gkoulalas-Divanis and colleagues[87]. Recent algorithms have focused on enhancing the utility of anonymized data [88-90] and applying anonymization to distributed data networks [91]. Anonymizing unstructured text is a far more difficult endeavor than structured data anonymization, similarly to data de-identification, but the impact on clinical information is potentially far more destructive. Chakaravarthy et al. [92] and Jiang et al. [93] have applied privacy models, the K-safety model for the former (prevents matching documents to entities based on terms that co-occur in a document), and t-plausibility for the latter (requires documents to be associated with at least t other plausible documents, any of which could be the original one, using word ontologies).

As discussed, de-identified data is often not anonymous, and the risk of re-identification, i.e. of linking a patient identity with de-identified data, can sometimes be important. For example, more than 96% of 2,700 patient records involved in a genome-wide association study were shown to be uniquely re-identifiable based on diagnosis codes [94]. However, the risk for patient re-identification in de-identified structured data sets has been assessed as low or very low. [95-98] Methods to estimate this risk with anonymized data sets were proposed by Dankar and colleagues [99]. Evaluating this risk for unstructured text has not been attempted using similar statistical approaches, but the empirical risk for a physician to recognize his patients in de-identified clinical notes was measured as very low [100].

D Data Integration, Interoperability, and Systems Federation

Data integration is an essential prerequisite in order to obtain clinical data from EHR systems. Current EHRs, depending on the clinical site, comprise up to 400-600 different IT systems which are networked using standards such as Health Level 7 (HL7) for textual data and Digital Imaging and Communications in Medicine (DICOM) for imaging data, often via commercial communication engines (e.g., eGate, Cloverleaf, or successors) [101, 102]. Integrating the Healthcare Enterprise (IHE) profiles, starting with clinical use cases, has successfully demonstrated how information transactions based on existing standards can be used to integrate the healthcare enterprise [103, 104].

Interestingly, most published data reuse projects do not use this type of horizontal data integration between operative quantity-based systems such as Patient Data Management Systems (PDMS), laboratory systems, Radiology Information Systems (RIS), and Picture archiving and communication systems (PACS). Instead, data reuse relies on vertical data integration which is typically reflected in data warehouse architectures, to be filled from source systems with copied data using an ETL (extraction-transformation-loading) process [105-107]. This approach is chosen because source data can thus be cleansed and filtered. Routine EHR data, for example, may comprise temporary data items, preliminary data items, and administrative data which are not desired within the research database. The process of copying data in a data warehouse architecture implies modification of both the source data structure and the data storage scheme. While routine EHR systems are transaction-oriented and must ensure data consistency when new data items are stored, extracted data in data warehouse structures is typically query-oriented. Instead of inserting single data items into the data warehouse, the ETL process will rather copy either the complete data source, or the delta since last import into the data warehouse. In addition, the ETL process supports the integration of data items from many different source systems as long as a common identifier such as a patient ID or case number can be used to join this data. Within the ETL process, it is typically possible to deal with missing data and data that does not fulfil consistency rules.

Data warehouse applications and ETL functionalities are available from many commercial vendors. For clinical data reuse however, it may be desirable to use open source toolsets to allow for cross-institutional data exchange. These tools offer several advantages such as unlimited access of many researchers in terms of licensing and the option for researchers to create their own specific queries, which is often limited in a commercial data warehouse environment. It can be observed that open source platforms such as i2b2 (Informatics for Integrating Biology and the Bedside) combined with open source ETL tools such as Talend Open Studio have been used in several data reuse projects [106-108]. Figure 1 depicts the architecture developed within the German Integrated Data Repository Toolkit (IDRT) to support integration of various operative source systems and different terminologies into an i2b2 research database.

Due to the privacy concerns mentioned above, the need for a scaled architecture may arise which ensures that local and pseudonymized data do not leave the source site. Such scaled architectures have been proposed e.g. within the EHR4CR project [109] to support the cooperation between local and central data warehouse structures using a so called “EHR4CR endpoint.” Thus, it is possible to support cohort selection of appropriate study patients across various sites and to collect patient informed consent only in a second step for the finally selected patients. Another technically interesting approach from the Scandinavian countries relies on the use of openEHR to extract data from several source EHRs [110].

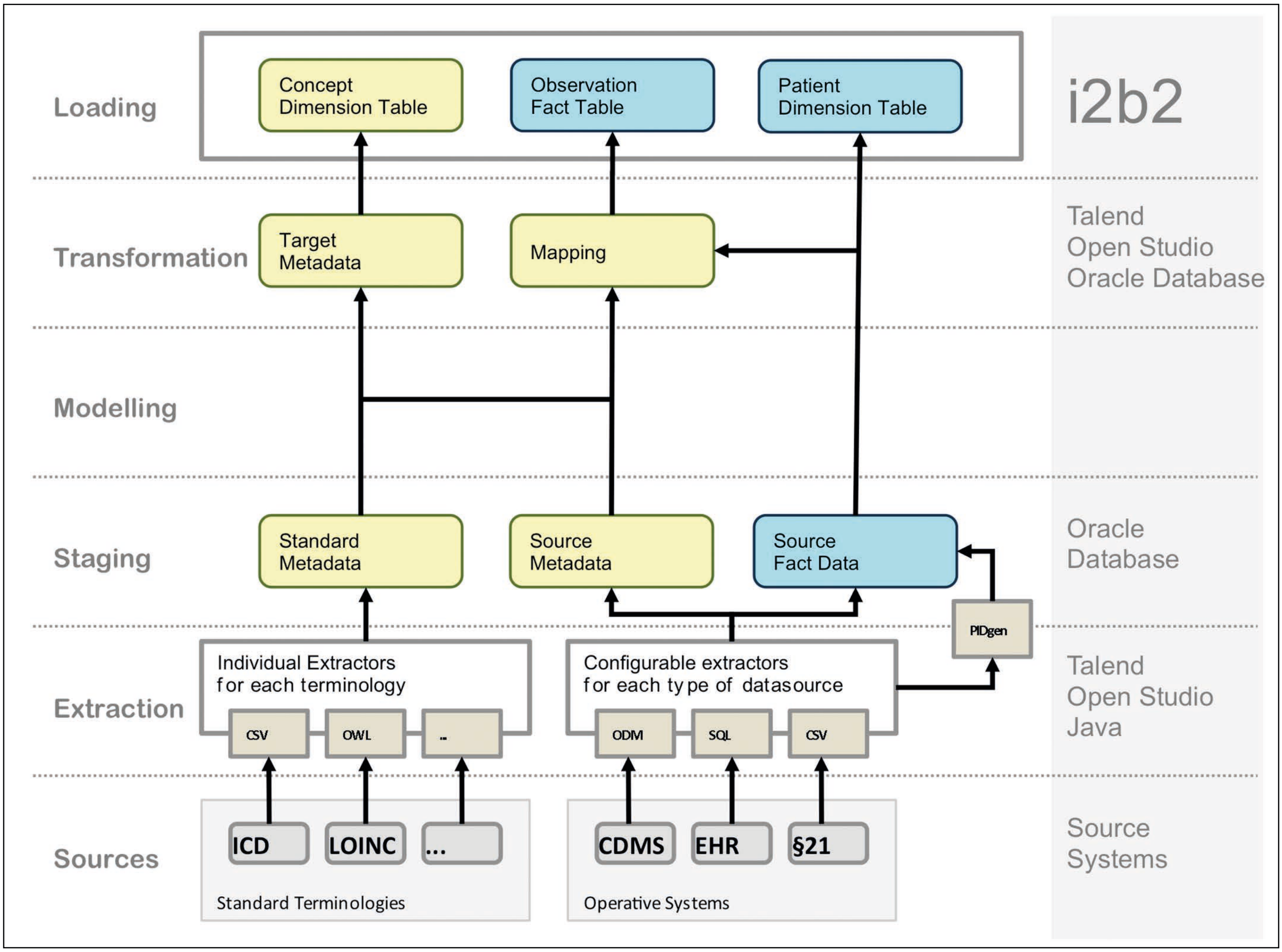

Fig. 1 Example of data extraction process from operative systems and source terminologies into an i2b2 research database infrastructure. Figure adapted from the IDRT project [107].

E Data Models and Terminologies Enabling Clinical Data Reuse

It has long been recognized that data transfer between different EHR systems relies on both syntactic and semantic constraints (Fig 2) [111, 112]. Data reuse projects face a similar problem. It is insufficient to simply transfer data into the research database without contextual knowhow of their meaning at that time. First generation interfaces used for EHR data transfer such as HL7 version 2.x covered the syntactic part of data transport only. In comparison, HL7 v3 defined a reference information model (RIM) to ensure a common understanding between the interfaced systems regarding transferred data contents. But its use has been hampered when existing EHR systems had different data models.

A powerful tool to improve semantic interoperability is the use of controlled terminologies [113]. Medicine has sought to ensure a common understanding by defining a growing number of classifications, nomenclatures, and ontologies such as the International Classification of Diseases (ICD) for diagnoses, the International Classification of Procedures in Medicine (ICPM) and many national procedure classifications, Logical Observation Identifiers Names and Codes (LOINC) for laboratory values, and the Systematized Nomenclature of Medicine (SNOMED) as an international nomenclature, to mention a few examples. Most medical terminologies have been developed for a specific purpose such as death statistics, health statistics, or billing. The use of terminologies for a common understanding of research data is essential to improve semantic interoperability. This can be seen in Figure 1 where the research database is constructed using such terminologies.

The Clinical Data Interchange Standards Consortium (CDISC [114]) is a non-profit organization developing standards for the exchange of digital clinical study data among associations. The principal software component within clinical studies is the electronic case report form (eCRF). An eCRF typically contains fields for data to be collected for one study subject according to the study protocol in a single clinical trial encounter [115]. There are many different options to structure a clinical trial, thus an electronic data capture (EDC) system must support a flexible definition of eCRFs. The CDISC consortium defined a set of standards for data capture, data transfer, and data analysis to facilitate data exchange between different study sites and their respective EDC systems. These standards include the XML-based Operational Data Model (ODM) to construct and model customized eCRF, and the Clinical Data Acquisition Standards Harmonization (CDASH) model, which defines the recommended data collection fields for 16 domains (version 1.1) such as patient demographics, concomitant medications, laboratory test results, or adverse events [116, 117].

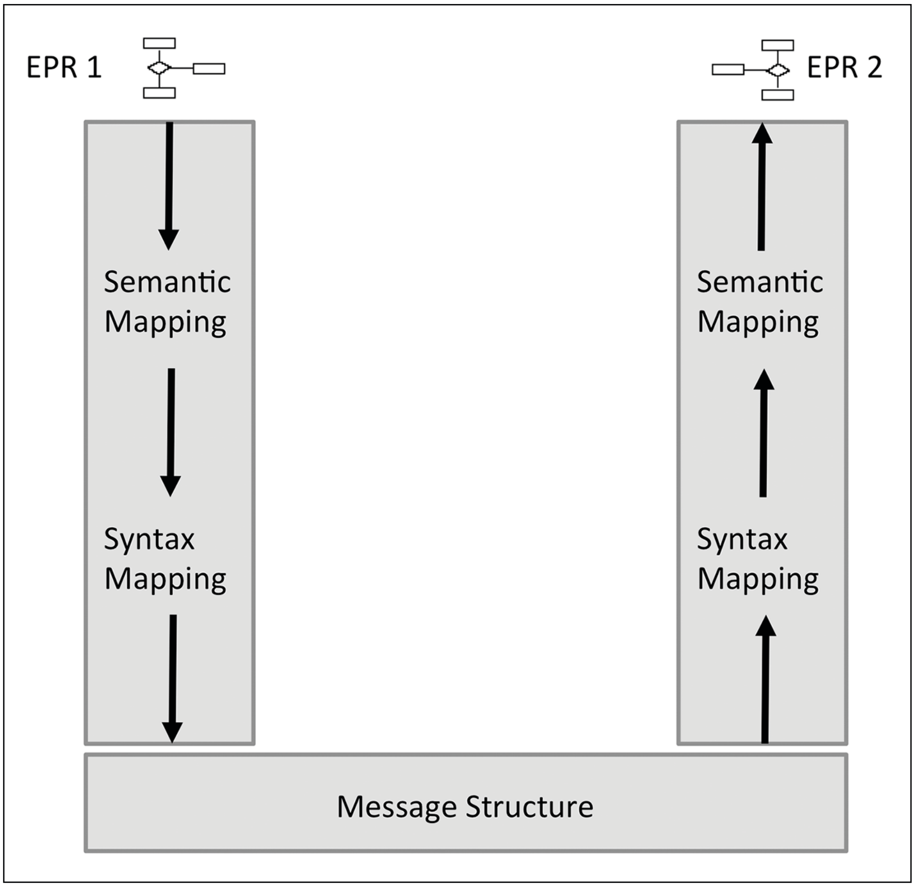

Fig. 2 Requirement for syntactic and semantic mapping when transferring data from one Electronic Patient Record (EPR) to another (adapted from [112]).

The following consequences arise for clinical data reuse: the research data warehouse should have an appropriate data scheme which maps source data during the ETL process to existing classifications and nomenclatures such as ICD, LOINC, Medical Dictionary for Regulatory Activities (MedDRA), Anatomical Therapeutic Chemical (ATC), or SNOMED. The Observational Health Data Sciences and Informatics (OHDSI) collaborative tries to force such mapping to common domain vocabularies [118, 119]. If data is to be reused for cohort identification only, this, in combination with the NLP methods mentioned in the following section could already be sufficient. The Patient-Centered Outcomes Research Institute (PCORI) has been launched in 2013 in the U.S. with a national Patient-Centered Clinical Research Network (PCORNet) to support interoperable clinical data research networks (CDRN) integrating patient-generated data and electronic health information for comparative effectiveness research [120,121]. For example, the New York City CDRN focuses on diabetes mellitus as common condition, and cystic fibrosis as rare condition [122].

F Extraction of Information from Unstructured Clinical Data

The majority of clinical information is stored in unstructured text format. In a recent survey of U.S. hospitals equipped with advanced EHRs, only about 35 % of their clinical data was captured in structured format, and 65% in unstructured text [123]. Reuse of this unstructured data requires either manual abstraction, or automated information extraction approaches based on NLP [124]. Most information extraction efforts focused on phenotyping and chart abstraction improvement [125], research subjects recruitment and cohort identification for retrospective studies, and patient identification for improved treatment and follow-up. The extraction of phenotypes and other types of information include diseases and problems, investigations, treatments, combined in the 4th i2b2 NLP challenge [126], or medication details for example [127]. Various data and attribute values were extracted to support peripheral artery disease and heart failure research in the eMERGE network [128], and to support obesity research [129]. Study subjects recruitment is a constant struggle, and adding more detailed information extracted from unstructured data to existing diagnostic codes significantly improves it [130]. Pakhomov and colleagues used it to identify patients suffering from angina pectoris [131] or heart failure [132]. Ni and colleagues used it to improve oncology trial eligibility screening [130], and Weng and Boland to represent and extract trial eligibility criteria [133, 134]. Extracting information to improve treatment and follow-up of patients has been applied to pancreatic [135] and colon neoplasms detection [136], thromboembolism and incidental findings [137], adverse events and errors detection [137], and patients acuity prediction [138]. Finally, information extracted from unstructured clinical data has been used to enable other examples of data reuse discussed below.

In several studies, NLP is used in combination with text- and data-mining. Typically, NLP is performed as the first processing step to extract medical concepts from narrative and unstructured portions of EHRs, while text- and data-mining techniques are applied to the data previously extracted with NLP. Some studies applied standard NLP techniques, such as cTAKES, MedLEE, and MetaMap, others applied ‘custom-made’ NLP techniques. Examples of the combined use of standard NLP and text- and data-mining are found in [139-141] where cTAKES is used with Boolean logic to perform phenotyping and to extract drug-side effects. MedLEE was applied for: 1) adverse drug reaction (ADR) signaling, where the association between a drug and an ADR was obtained by using disproportionality analysis [142, 143] or Boolean logic [144], or by building and analyzing statistical distributions of concepts (i.e., diseases, symptoms, medications) extracted from the narrative text [145]; 2) EHR-data driven phenotyping using Boolean logic on MedLEE-extracted concepts [136, 146]; 3) automated classification of outcomes from the analysis of emergency department computed tomography imaging reports using machine learning methods, such as decision trees [147]. MetaMap has been used with logistic regression in [148] to discover inappropriate use of emergency room based on information on drugs, psychological characteristics, diagnoses, and symptoms. Finally, a review of the application of standard NLP methods combined with data mining can be found in [149].

In other cases, NLP is implemented using basic text search of a list of ‘key words’ identified by the authors and subsequent analysis of the set of terms extracted with Boolean logic [150,151], disproportionality analysis [152], contingency tables,[153] logistic regression [154], and classification methods [155]. Fields of applications include EHR-data driven phenotyping, ADR signaling, and the assessment of effects of mood instability on clinical outcomes. Finally, an example of use of ‘custom-made’ NLP systems is given in [156] where a NLP tool based on the French medical lexicon and UMLS is used with Boolean logic to analyze medical reports and automatically detect surgical site infections in neurosurgery.

G Mining Structured Clinical Data

The following is a brief description of the rationale and typical methods used for EHR data mining. Methods are clustered in 10 categories as discussed below.

Boolean logic extracts data using queries made by Boolean combinations of a set of conditions. Boolean logic was applied in many studies, i.e., [157] and [158], ranging from the analysis of EHRs for the evaluation of the effectiveness of triage models used in mass casualty research to the identification of emergent endotracheal intubation in ICU patients.

Fuzzy logic is used to solve problems where it is more convenient to consider the concept of ‘partial truth’: a variable might be partially true or partially false. An example is given in [159] where EHRs are analyzed to detect potential ADR signals.

Regression analysis models the relationships between a dependent variable and one or more independent variables. In logistic regression, the relationship between the dependent and the independent variable(s) is modeled with a cumulative logistic distribution. This method has been applied to predict crush syndrome from a set of risk factors,[160] to improve the performance of severity of illness scores [161], to model factors associated with overweight and obesity [162], to characterize differences in co-morbid profiles between different cohorts [163], to determine the association between nurse continuity and hospital-acquired pressure ulcers [164], to discover how the patient and the characteristics of support and intervention systems affect the improvement in urinary and bowel incontinence [165], and, finally, to detect ADR signals from EHRs [166]. In orthogonal regression, the relationship between the dependent and the independent variable(s) is the one that minimizes the orthogonal distances from the observed values of the dependent variable and the corresponding values on the fitting line. Sun and colleagues used orthogonal regression to identify risk factors related to an adverse condition[167].

The Apriori algorithm is the most widely known association rule algorithm using an iterative approach to find the most frequent associations between two or more items and gives a measure of the frequency with which that particular association has been found. The algorithm has been applied in [168] to discover associations between diagnoses of different sub-groups of patients. Association rule mining has been applied in

[169] to identify the associations between combination of diagnoses, demographics, and lab results to predict high risk of diabetes. In [170] association rule was applied to discover medical correlations, characterize data trends, and perform predictive analysis on data trends and medical correlations.

Classification is the process of assigning a new observation to a specific pre-defined category or class. In decision tree classification, a decision tree is used to predict the value of a target variable (or item) based on the observations of several input variables. Classification And Regression Tree (CART) analysis, a particular type of decision tree, has been applied to detect ADRs [171, 172]. The k-Nearest Neighbors (k-NN) algorithm, another classification method, assigns an object to the most common class among its k nearest neighbors. k-NN is used in [173] for retrieving patients with similar characteristics by analyzing EHRs. Fuzzy neural networks are the combination of neural networks and fuzzy logic. Skevofilakas and colleagues used fuzzy neural networks to predict the risk of Type I Diabetes Mellitus patients to develop diabetic retinopathy [174]. Finally, Support Vector Machines (SVM) aim at assigning a new observation into one of two possible categories. It was applied in combination with Bayesian networks and k-NN in [175] to predict pancreatic cancer.

Clustering aims at finding hidden patterns - the clusters - in a data set. In fuzzy-clustering, data are assigned to more than one cluster and are associated to a set of membership levels corresponding to the strength of the association between that data element and a particular cluster. In [176], fuzzy-clustering is used for the identification of rare-cases in post-operative pain management. Hierarchical clustering builds a hierarchy of clusters to find which clusters should be combined/agglomerated and which should be split or divided. In addition to [176], hierarchical clustering has been applied in [177] to identify periodic/seasonal patterns in incidence of diseases. Non-negative tensor factorization (NTF) is a technique to decompose large dimension data tensors containing non-negative elements as a product of two non-negative tensors of smaller size. Ho and colleagues applied NTF for EHR data-driven phenotyping based on the interaction between diagnoses and medications [178].

Relational data mining is the application of data mining techniques to relational databases. Chen and colleagues described the application of relational data mining to detect anomalies in the accesses to communities information systems [179]. The study by Peissig and colleagues used Inductive Logic Programming (ILP) - a method that infers an hypothesis from the analysis of the background knowledge and examples - to derive phenotypes from EHR data [180].

Disproportionality analysis (DPA) is a method typically used in the investigation of ADR signals. The information component, one of the most common DPA methods, measures the disproportionality between the association of two variables, such as a drug and an ADR, as in a study by Norén and colleagues[181].

Probabilistic graphical models, such as Bayesian networks, are a widely used class of structured prediction models. Graphic models describe the underlying relations between the variables with a graph: the links between the different variables represent the conditional dependencies between the variables. Bayesian networks together with k-NN and SVM were used in [175] to predict pancreatic cancer by using knowledge-base from PubMed research papers and experimental observations derived from EHRs. Graphic modeling is found also in [182] to identify which user accesses to EHR data deviate from the accesses found during typical patient care.

Topic modeling relies on statistical models for extracting the "topics" that occur in a set of documents. One of the models used in topic modeling is the Latent Dirichlet Allocation (LDA) where the statistical information is assumed to have a Dirichlet distribution. LDA was used in [183] for EHR-driven phenotyping and in [184] to discover which user accesses to EMR data differ from the typical access pattern.

Finally, some studies applied simultaneously multiple data mining methods, such as in [185] where different approaches ranging from disproportionality analysis to logistic regression are compared and used to detect ADR signals from EHRs. In [186], knowledge-base is used for EHR data-driven phenotyping for gene-disease association finding.

H Clinical Practice and Research Integration

While there are huge expectations at reusing data produced during care processes, there are also important challenges. Clinical documentation is a paramount activity of clinicians to track patient's conditions and communicate with other health professionals. However, measures to progressively improve and increase secondary usage of clinical data, from billing to quality assessment or from clinical research to public health, have increased purposes beyond the direct care of the patient. This has led to an important increased workload for care professionals [187]. Clinical documentation requires 25-50% of clinicians' time and, in a recent narrative review by Clynch and Kellett, there has been almost no formal research to assess its value, or on whether the time spent on it has negative effects on patient care [188]. There are now numerous reports about information and alerts overload using EHRs and its consequences [189, 190].

The integration of clinical practice and research can be considered from three major points of view: clinical practice to leverage clinical research, support for bedside clinical research, and data reuse to improve clinical practice.

For clinical practice to leverage clinical research, using common semantics is a major challenge. There are numerous publications and works that have tried to leverage clinical research in reusing data directly extracted from care records. This challenge is getting even more important with the increasing need of precise phenotype information for genomics and personalized medicine. Unfortunately, the lack of definition for phenotype descriptions has led to the proliferation of numerous definitions for most phenotypic information, including problems, patient history, physical examinations, conditions, and clinical profiles in general, among researchers, care providers, and for administration requirements. For example, Gregg and colleagues have reported in 2014 that the prevalence of some important complications of diabetes, such as neuropathy, chronic kidney disease, peripheral vascular disease, could not properly be assessed due to inconsistent EHR documentation and definitions across the United States for the 1990-2010 period [191]. There is still a lot of literature about addressing the challenge of unified semantics. Two different trends can be seen. The first trend is going towards semantic-centered EHR rather than data-centered systems, such as developing EHR systems based on openEHR [192-194] or robust semantic encoding using semantic rich resources, such as SNOMED [195]. However, both approaches remain relatively marginal and resource intensive, though they most probably offer the better perspectives. The second trend consists in bridging the EHR with external analytical tools through a complex ETL process that involves both data normalization and semantic alignment. Most systems available today, either in research & developement such as EHR4CR [11, 196], DebugIT [197], and i2b2, or as commercial products are based on such types of bridges. An important challenge is about the nature of data. For numerous reasons, EHRs tend to increase the amount of data. On one side, there is a strong push towards increasing the structuration of patient records. Structured data have a lot of nice characteristics, most of them can be re-used for decision-support in direct care, but also for numerous secondary usages. On the other side, need for speed and efficiency promotes (semi)-automatic production of documents, such as summaries, discharge documents, reports, and progress notes. When automatically processed, new documents are usually built from "copy-pasted" part of documents already existing in the patient record, thus increasing the volume of data without increasing the quantity of information [198].

Bedside clinical research is an important pillar of research in life sciences and the widespread adoption of EHRs provides a new opportunity to improve the efficiency of clinical research. However, the clinical research made "on a daily and pervasive" manner tends to be difficult for clinicians, mostly due to the pressure of efficiency and to the increasing number of requirements needed for clinical research. Providing efficient tools for clinicians to support their own clinical research, to build cooperative and collaborative networks of clinical researchers beyond the border of academic settings, and to do research in real settings, are major goals to be achieved. There are many initiatives that try to address these challenges, such as i2b2 in the U.S. [199] or EHR4CR in Europe [200]. Clinicians have been early adopters of EHRs to support their own clinical research, including in clinical practices [201]. However, this tends to be less the case, probably because of the reasons discussed above: efficiency pressure, overload of information, and higher requirements for clinical research.

How can data reuse improve clinical practice? Data is a major asset that should be considered as strategic for any clinical organization. This implies, for example, that data should never be only available in a legacy, proprietary repository. Data must be available under the full control of the organization with all the metadata required to allow data processing and analytics. One of the reasons for this is that clinical data of an organization behaves like a local and progressive knowledge about the presentation, conditions, and evolution of patients specific to this organization, considering the prevalence of presentation and conditions of this cohort of patients, in relation with the care and means available in the organization. It allows to implement the paradigm quoted by Ilias Iakovidis "Medicine is a global science and a local art."[202] There are several ways data reuse can improve clinical practice: 1) Improve the patient record and decision support: this is the reuse of data within the same patient record, avoiding duplicates, connecting data, supporting inferences and decision-support, coupling knowledge with external sources of information, amongst others; 2) Cases/peers comparison for a continuous learning process: cases and peer comparison could be a much more powerful instrument in EHR. It can be used in real-time and has been shown to be effective by several authors, i.e., Milchak and colleagues [203]. 3) Build contextualized case-based database and improve the predictive values of decision support: most EHRs implement decision support in various forms, however they rarely consider the prevalence of conditions used in decision support. Predictive values, especially the positive predictive value in the case of CPOEs, is closely linked to the prevalence of the alert considered. This has been demonstrated for drug-drug interactions decision support that has a very low positive predictive value [204, 205]. Using the characteristics of the local population of patients of a given organization can provide precise and real-time prevalence, thus allowing to adapt decision support and improve its positive predictive value. Data-driven approaches using large datasets have also been tested, e.g., for computing risk factors [206]. 4) Engage patients: this point is now receiving a large audience with the Blue Button initiative, that allows patients to access, or download, their own patient record [207].

Clinical Data Reuse Examples

a) Quality measurements extraction: Clinical Quality Measures (CQMs) are used for assessing processes, access, outcomes, structure, experience, management, or efficiency of patient care. As defined by the U.S. Centers for Medicare & Medicaid Services (CMS), CQMs assess "the degree to which a provider competently and safely delivers clinical services that are appropriate for the patient in an optimal timeframe."[208] The CMS Quality Measures Inventory [209] lists more than 1,500 measures (in February 2016), and the National Quality Measures Clearinghouse (NCQM [210]) more than 2,100 (in February 2016). Among these measures, about 400 are endorsed by the National Quality Forum (NQF [211]). Several CQMs are required by the U.S. Medicare and Medicaid incentive program to demonstrate "meaningful use" of EHRs. The automatic extraction of CQMs from clinical notes has been attempted with only a few clinical note types (e.g., colonoscopy reports) or disease categories (e.g., heart failure). Examples focused on colonoscopy reports included assessing the reports' quality [212], and detecting patients with polyps or adenomas. Gawron and colleagues developed a NLP application reaching 94% recall and precision when detecting the location and histology of adenomas, and 69% when counting their number [213]. Raju and colleagues compared a manual abstraction with an NLP-based process to extract screening information, correctly identifying 91.3% of them with NLP, and 87.8% manually [214]. Studies focused on heart failure targeted the extraction of mentions and values of left ventricular ejection fraction [24], a key functional test for assessing heart failure, and added heart failure treatment information to functional testing to automatically detect patients not treated according to published recommendations. The latter study was based on the Congestive Heart failure Information Extraction Framework (CHIEF), an application based on NLP to automatically extract left ventricular functional testing results [215, 216], heart failure treatment medications [217], and reasons not to prescribe these medications, eventually detecting patients not treated according to recommendations with 98.9% sensitivity, and 98.7% positive predictive value [218].

b) Learning healthcare systems: The concept of Learning Healthcare System (LHS), defined by the American Institute of Medicine (IOM) in 2007 is emerging as a perfect example of clinical data reuse stimulating improvement of healthcare services. LHS is often characterized as a continuous loop of health data collection, knowledge extraction and its application in clinical practice, which starts a new iteration of the LHS [219]. Fast progression of knowledge into health service delivery, improved adaptation to individual patient needs, and support for shared clinical decision-making are highlighted as major advantages originating from health data reuse.

A review of activities transforming healthcare services into agile and adaptive learning systems highlighted a relatively low success rate currently reflected in literature. Even though the interest on exploring the ideas of LHS is global, implementations in practice are few [220]. Many initiatives including several IOM meeting reports focus on conceptual challenges hindering the adoption of LHS [221-223]. Getting access to EHR data and making use of structured and unstructured information trigger an avalanche of problems without a straightforward solution. Development of comprehensive data models enabling semantic interoperability of data accumulated in various healthcare systems is pursued by many research groups [224-226], promising a solid foundation for clinical data reuse (as discussed in more details in sections E and F). However, much research is still needed to turn these ideas into reality.

Regardless of many challenges, several research initiatives managed to demonstrate the principles of the LHS in practice. The scale of reported studies varies from hundreds [227] to millions [228] individual patient records processed by distributed or centralized infrastructures. EHR data is often combined with patient reported outcomes to better address the aims of the LHS paradigm [220]. It provides a better understanding of "patient data shadow" [229] enabling personalization of care. The aforementioned projects suggest that health data can and will be used for improving the performance and quality of healthcare, lowering costs, and addressing the individual needs of the patient to a larger extent in the future.

While successful implementations of LHS are reported, their impact remains poorly documented [220]. The benefits for patients, health services, and society are difficult to measure, however, knowing them could lead to faster adoption of data reuse practices and improve their acceptance by healthcare professionals. Currently, much effort is directed towards succeeding in technology development (semantic interoperability, data access, and processing mechanisms), while mapping this effort to the aims of a modern healthcare (improved patient care experience, better population health, and reduced costs) often remains unclear [230].

IV Discussion and Potential Future Progress

As explained earlier, reuse of clinical data is crucial for healthcare quality, management, reduced costs, population health management, and effective clinical research. This need has been widely recognized and numerous efforts have been reported in the scientific literature and included in this review.

As limitations, this review only includes the works reported in scientific publications and focuses on a selection of some aspects of clinical data reuse that were considered important by the authors. It was not intended to do a comprehensive review of all published works in this field. Only a selection of bibliographic databases was used (MEDLINE, Web of Science, and conference proceedings). We used a conceptual model of clinical data reuse that was developed for this review only. This conceptual model is partly reflected in the sections included in this review. Legal and policy issues framing clinical data reuse and examples of clinical data reuse (clinical research, clinical research subject recruitment, public health surveillance) are some additional important aspects of clinical data reuse that were included in the conceptual model but not in the final review.

In a recent review focused on the reuse of structured data, Vuokko and colleagues found that most publications report how clinical data reuse should impact care processes, productivity and costs, patient safety, care quality, or health outcomes, rather than what actual studies did realize when reusing clinical data [231]. Most research demonstrating each of these possible advantages of clinical data reuse still lies in our future.

Opportunities for future progress are numerous, ranging from new legislations easing clinical data reuse while protecting patient privacy, to the addition of other types of observational data (e.g., consumer-provided data, personal and quantified self sensor data, genomic and microbiota data, environment data), and larger-scale applications. As a good example of the latter, the OHDSI collaborative [118] growing infrastructure is making very large scale studies based on reused observational data, potentially including hundreds of millions or even billions of research subjects!

Acknowledgments
We would like to thank Hans-Ulrich Prokosch and Johan Gustav Bellika for their initial contributions to this review.

References

1. U.S. Department of Health and Human Services. Doctors and hospitals' use of health IT more than doubles since 2012. Available from: http://www.hhs.gov/news/press/2013pres/05/20130522a.html
2. Schoen C, Osborn R, Doty MM, Squires D, Peugh J, Applebaum S. A survey of primary care physicians in eleven countries, 2009: perspectives on care, costs, and experiences. Health Affairs Project HOPE - The People-to-People Health Foundation, Inc; 2009 Nov;28(6):w1171–83.
3. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007 Jan;14(1):1–9.
4. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington (DC): National Academies Press (US); 2011.
5. National Committee on Vital and Health Statistics. Information for health: A strategy for building the National Health Information Infrastructure. Available from: aspe.hhs.gov/sp/nhii/documents/NHIIReport2001/toc.htm
6. Fries JF, McShane DJ. ARAMIS (the American Rheumatism Association Medical Information System). A prototypical national chronic-disease data bank. West J Med 1986 Dec;145(6):798–804.
7. Centers for Medicare Medicaid Services. 42 CFR 424: Medicare Program; Electronic Submission of Medicare Claims. Available from: https://www.federalregister.gov/articles/2005/11/25/05-23080/medicare-program-electronic-submission-of-medicare-claims
8. Charles D, Gabriel M, Furukawa M. Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008-2013. Available from: http://www.healthit.gov/sites/default/files/oncdatabrief16.pdf
9. Lehmann CU, O'Connor KG, Shorte VA, Johnson TD. Use of electronic health record systems by office-based pediatricians. Pediatrics 2015 Jan;135(1):e7–15.
10. Payne TH, Detmer DE, Wyatt JC, Buchan IE. National-scale clinical information exchange in the United Kingdom: lessons for the United States. J Am Med Inform Assoc 2011 Jan;18(1):91–98.
11. Fadly El A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P-Y, et al. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011 Dec;44 Suppl 1:S94–102.
12. Steele SJ. Working with the CTSA Consortium: what we bring to the table. Sci Transl Med 2010 Dec 22;2(63):63mr5–63mr5.
13. Rea S, Pathak J, Savova G, Oniki TA, Westberg Les, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform 2012 Aug;45(4):763–71.
14. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011;4(1):13.
15. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform 2014;9:97–104.
16. Gartner Inc. What Is Big Data? - Gartner IT Glossary - Big Data. 2012. Available from: http://www.gartner.com/it-glossary/big-data/
17. Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes: Healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis. J AHIMA 2012 Oct;83(10):38–43– quiz 44.
18. Safran C. Using routinely collected data for clinical research. Stat Med 1991 Apr;10(4):559–64.
19. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care 2007 Jun;13(6 Part 1):277–8.
20. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med 2009 Sep;151(5):359–60.
21. Safran C. Reuse of Clinical Data. Yearb Med Inform 2014;9(1):52–4.
22. Price Waterhouse Coopers. Transforming Healthcare Through Secondary Use of Health Data; 2009.
23. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):48–855.
24. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc 2012 Aug 9;19(5):859–66.
25. Herrmann FR, Safran C, Levkoff SE, Minaker KL. Serum albumin level on admission as a predictor of death, length of stay, and readmission. Arch Intern Med 1992 Jan;152(1):125–30.
26. McManus DD, Saczynski JS, Lessard D, Waring ME, Allison J, Parish DC, et al. Reliability of Predicting Early Hospital Readmission After Discharge for an Acute Coronary Syndrome Using Claims-Based Data. Am J Cardiol 2016 Feb 15;117(4):501-7.
27. Cai X, Perez-Concha O, Coiera E, Martin-Sanchez F, Day R, Roffe D, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. J Am Med Inform Assoc 2015 Sep 15;:ocv110.
28. Temple MW, Lehmann CU, Fabbri D. Predicting Discharge Dates From the NICU Using Progress Note Data. Pediatrics 2015 Aug;136(2):e395–405.
29. Evans RS, Burke JP, Classen DC, Gardner RM, Menlove RL, Goodrich KM, et al. Computerized identification of patients at high risk for hospital-acquired infection. Am J Infect Control 1992 Feb;20(1):4–10.
30. Pittet D, Safran E, Harbarth S, Borst F, Copin P, Rohner P, et al. Automatic alerts for methicillin-resistant Staphylococcus aureus surveillance and control: role of a hospital information system. Infect Control Hosp Epidemiol 1996 Aug;17(8):496–502.
31. Samore MH, Lichtenberg D, Saubermann L, Kawachi C, Carmeli Y. A clinical data repository enhances hospital infection control. Proc AMIA Annu Fall Symp 1997;:56–60.
32. Wagner MM, Moore AW, Aryel RM. Handbook of biosurveillance; 2011.
33. Kopcke F, Prokosch H-U. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014;16(7):e161.
34. Weiner MG, Xie D, Tannen RL. Clinical trials in silico: rigorous assessment of treatment effect using electronic health records. AMIA Annu Symp Proc 2008;:1172.
35. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013 Oct 18;274(6):547–60.
36. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29.
37. Harper EM. The economic value of health care data. Nurs Adm Q 2013 Apr;37(2):105–8.
38. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2012 Sep 6.
39. Overhage JM, Overhage LM. Sensible use of observational clinical data. Stat Methods Med Res 2013 Feb;22(1):7–13.
40. Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. J Law Med Ethics. 2013 Mar;41 Suppl 1(s1):56–60. PMID: 23590742
41. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. J Biomed Discov Collab 2011;6:48–52. PMID: 21647858
42. Ancker JS, Shih S, Singh MP, Snyder A, Edwards A, Kaushal R. Root causes underlying challenges to secondary use of data. AMIA Annu Symp Proc 2011;2011:57–62.
43. de Lusignan S, Pearce C, Shaw NT, Liaw S-T, Michalakidis G, Vicente MT, et al. What are the barriers to conducting international research using routinely collected primary care data? Stud Health Technol Inform 2011;165:135–40.
44. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Summits Transl Sci Proc 2010;2010:1–5.
45. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003 May 17;326(7398):1070.
46. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc 2013;2013:1472–7.
47. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care 2012 Jul;50 Suppl:S21–9.
48. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed

Inform 2013 Oct;46(5):830–6.
49. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan;20(1):144–51.
50. Silversides A. Privacy concerns raised over "secondary use" of health records. CMAJ 2009 Dec;181(12):E287.
51. Emam El K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015;350(mar20 1):h1139.
52. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform 2013 Jan;82(1):1–9.
53. Hersh WR, Cimino J, Payne PRO, Embi P, Logan J, Weiner M, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. EGEMS (Wash DC) 2013;1(1):1018.
54. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013 Aug;51(8 Suppl 3):S30–7.
55. Health Information & the Law. Who Owns Medical Records: 50 State Comparison. Available from: http://www.healthinfolaw.org/comparative-analysis/who-owns-medical-records-50-state-comparison
56. Shiyanbola OO, Mort JR. Patients' perceived value of pharmacy quality measures: a mixed-methods study. BMJ Open 2015 Jan 19;5(1):e006086.
57. Ganguli I. Stuck on loop: why do patients have to repeat their stories? Available from: http://healthydebate.ca/opinions/stuck-on-loop-why-do-patients-have-to-repeat-their-stories
58. International Medical Informatics Association. The IMIA Code of Ethics for Health Information Professionals. Available from: http://www.imia-medinfo.org/new2/pubdocs/Ethics_Eng.pdf
59. GPO US. CFR Title 45 Subtitle A Part 164: Security and Privacy. Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html
60. GPO US. CFR Title 45 § 46: Protection of Human Subjects. Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html
61. Directive 95/46/EC of the European Parliament and of the Council. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
62. United Kingdom. Data protection act 1998. Available from: http://legislation.gov.uk/ukpga/1998/29
63. Iacono Lo L. Multi-centric universal pseudonymisation for secondary use of the EHR. Stud Health Technol Inform 2007;126:239–47.
64. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. Med Care 2012 Jul;50 Suppl:S82–S101.
65. Somolinos R, Munoz A, Elena Hernando M, Pascual M, Caceres J, Sanchez-de-Madariaga R, et al. Service for the Pseudonymization of Electronic Healthcare Records Based on ISO/EN 13606 for the Secondary Use of Information. IEEE J Biomed Health Inform 2015 Nov;19(6):1937–44.
66. Chan EM, Lam PE, Mitchell JC. Understanding the Challenges with Medical Data Segmentation for Privacy. HealthTech'13 Washington, DS; 2013. p. 1–10.
67. Loukides G, Liagouris J, Gkoulalas-Divanis A, Terrovitis M. Disassociation for electronic health record privacy. J Biomed Inform 2014 Aug 1;50(C):46–61.
68. Meystre SM. De-identification of Unstructured Clinical Data for Patient Privacy Protection. Medical Data Privacy Handbook; 2015. p. 697–716.
69. Dorr DA, Phillips WF, Phansalkar S, Sims SA, Hurdle JF. Assessing the difficulty and time cost of de-identification in clinical narratives. Methods Inf Med 2006;45(3):246–52.
70. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10:70.
71. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating current automatic de-identification methods with Veteran's health administration clinical documents. BMC Med Res Methodol 2012 Jul 27;12(1):109.
72. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Generalizability and comparison of automatic clinical text de-identification methods and resources. AMIA Annu Symp Proc 2012;2012:199–208.
73. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010 Dec;79(12):849–59.
74. Ferrandez O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013 Jan 1;20(1):77–83.
75. Zuccon G, Kotzur D, Nguyen A, Bergheim A. De-identification of health records using Anonym: Effectiveness and robustness across datasets. Artif Intell Med 2014 Jul;61(3):145–51.
76. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007 Sep;14(5):550–63.
77. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform 2015 Sep 26;:1–9.
78. Meystre SM, Ferrandez O, Friedlin FJ, South BR, Shen S, Samore MH. Text de-identification for privacy protection: a study of its impact on clinical text information content. J Biomed Inform 2014 Aug;50:142–50.
79. Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc 2013 Mar;20(2):342–8.
80. McMurry AJ, Fitch B, Savova G, Kohane IS, Reis BY. Improved de-identification of physician notes through integrative modeling of both public and private medical text. BMC Med Inform Decis Mak 2013;13(1):112.
81. Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform 2014 Apr;83(4):303-12.
82. Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform 2014 Aug 1;50(C):151–61.
83. Dalianis H, Velupillai S. De-identifying Swedish clinical text - refinement of a gold standard and experiments with Conditional random fields. J Biomed Sem 2010;1(1):6.
84. Sweeney L. K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 2002 Oct;10(05):557–70.
85. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-Diversity: Privacy Beyond k-Anonymity. ICDE IEEE; 2006;:24.
86. Dankar FK, Emam El K. Practicing Differential Privacy in Health Care: A Review. TDP 2013;6(1):35–67.
87. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. J Biomed Inform 2014 Aug 1;50(C):4–19.
88. Tamersoy A, Loukides G, Nergiz ME, Saygin Y, Malin B. Anonymization of longitudinal electronic medical records. IEEE Trans Inf Technol Biomed 2012 May;16(3):413–23.
89. Loukides G, Gkoulalas-Divanis A. Utility-aware anonymization of diagnosis codes. IEEE journal of biomedical and health informatics IEEE J Biomed Health Inform; 2013 Jan;17(1):60–70.
90. Xia W, Heatherly R, Ding X, Li J, Malin BA. R-U policy frontiers for health data de-identification. J Am Med Inform Assoc 2015 Sep;22(5):1029–41.
91. Kohlmayer F, Prasser F, Eckert C, Kuhn KA. A flexible approach to distributed data anonymization. J Biomed Inform 2013 Dec 12.
92. Chakaravarthy VT, Gupta H, Roy P, Mohania MK. Efficient techniques for document sanitization. CIKM New York, New York, USA: ACM Press; 2008;:843–852.
93. Jiang W, Murugesan M, Clifton C, Si L. t-Plausibility: Semantic Preserving Text Sanitization. CSE IEEE; 2009;3:68–75.
94. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc 2010 May;17(3):322–7.
95. Emam El K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. J Med Internet Res 2006;8(4):e28.
96. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. J Am Med Inform Assoc 2010 Feb 26;17(2):169–77.
97. Kwok PK, Lafky D. Harder Than You Think: A Case Study of Re-Identification Risk of HIPAA-Compliant Records. Joint Statistical Meetings 2011 2011. p. 1–8.
98. Emam El K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071.
99. Dankar FK, Emam El K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak 2012 Jul 9;12(1):1–1.
100. Meystre S, Shen S, Hofmann D, Gundlapalli A. Can Physicians Recognize Their Own Patients in De-identified Notes? Stud Health Technol Inform

2014;205:778–2.

101. Heitmann KU. The role of communication servers in the architecture of healthcare information systems. Stud Health Technol Inform 1997;45:156–62.

102. Wentz B, Seggewies C, Bell R, Knispel S, Kraska D. The Erlangen Hospital Communication Hub: migration from proprietary to standardised communication. Stud Health Technol Inform 1997;45:163–67.

103. Channin DS. Integrating the Healthcare Enterprise: a primer. Part 2. Seven brides for seven brothers: the IHE integration profiles. Radiographics 2001 Sep;21(5):1343–50.

104. Bernardini A, Alonzi M, Campioni P, Vecchioli A, Marano P. IHE: integrating the healthcare enterprise, towards complete integration of healthcare information systems. Rays 2003 Jan;28(1):83–93.

105. Bichutskiy VY, Colman R, Brachmann RK, Lathrop RH. Heterogeneous biomedical database integration using a hybrid strategy: a p53 cancer research database. Cancer Inform 2007 Feb 20;2:277–87.

106. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform 2010;160(Pt 1):193–7.

107. Bauer CRKD, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M, et al. Integrated Data Repository Toolkit (IDRT). A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data. Methods Inf Med 2016;55(2):125–35.

108. Ganslandt T, Mate S, Helbing K, Sax U, Prokosch HU. Unlocking Data for Clinical Research - The German i2b2 Experience. Appl Clin Inform 2011;2(1):116–27. PMID: 23616864

109. Fadly El A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P-Y, et al. Integrating clinical research with the Healthcare Enterprise. J Biomed Inform 2011 Dec;44(S1):S94–S102.

110. Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG. Archetype-based data warehouse environment to enable the reuse of electronic health record data. Int J Med Inform 2015 Sep;84(9):702–14.

111. Schweiger R, Bürkle T, Hölzer S, Tafazzoli AG, Dudeck J. Plug and play--fiction or reality? Stud Health Technol Inform 1998;52 Pt 2:999–1001.

112. Bürkle T, Schweiger R, Altmann U, Holena M, Blobel B, Dudeck J. Transferring data from one EPR to another: content--syntax--semantic. Methods Inf Med 1999 Dec;38(4-5):321–5.

113. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. AMIA Annu Symp Proc 2013;2013:648–56.

114. Kuchinke W, Wiegelmann S, Verplancke P, Ohmann C. Extended cooperation in clinical studies through exchange of CDISC metadata between different study software solutions. Methods Inf Med 2006;45(4):441-6.

115. Mathura VS, Rangareddy M, Gupta P, Mullan M. CliniProteus: A flexible clinical trials information management system. Bioinformation 2007;2(4):163–5.

116. Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med 2009;48(5):408–13.

117. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. Methods Inf Med 2009;48(1):45–54.

118. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI) - Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574–8.

119. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, et al. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. EGEMS (Wash DC) 2016;4(1):1189.

120. Selby JV, Krumholz HM, Kuntz RE, Collins FS. Network news: powering clinical research. Sci Transl Med 2013 Apr 24;:182fs13.

121. PCORnet PPRN Consortium, Daugherty SE, Wahba S, Fleurence R. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. J Am Med Inform Assoc 2014 Jul;21(4):583-6.

122. Kaushal R, Hripcsak G, Ascheim DD, Bloom T, Campion TR, Caplan AL, et al. Changing the research landscape: the New York City Clinical Data Research Network. J Am Med Inform Assoc 2014 Jul;21(4):587–90.

123. Cannon J, Lucci S. Transcription and EHRs: Benefits of a Blended Approach. Journal of AHIMA 2010 Feb;81(2):36–40. Available from: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_046429.hcsp?dDocName=bok1_046429

124. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;:128–44.

125. Carrell DS, Halgrim S, Tran D-T, Buist DSM, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014 Mar 15;179(6):749–58.

126. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Aug 16;18(5):552–6.

127. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010 Jan;17(1):19–24.

128. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Summits Transl Sci Proc 2013;2013:149–53.

129. Kreuzthaler M, Schulz S, Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. J Biomed Inform 2015 Feb 1;53(C):188–95.

130. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility Pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak 2015;15(1):28.

131. Pakhomov SSV, Hemingway H, Weston SA, Jacobsen SJ, Rodeheffer R, Roger VL. Epidemiology of angina pectoris: role of natural language processing of the medical record. Am Heart J 2007 Apr;153(4):666–73.

132. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007 Jun;13(6 Part 1):281–8.

133. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011 Dec;18 Suppl 1(Supplement 1):i116–24.

134. Boland MR, Tu SW, Carini S, Sim I, Weng C. EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria. AMIA Summits Transl Sci Proc 2012;2012:71–80.

135. Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. HPB (Oxford) 2010 Dec;12(10):688–95.

136. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011;2011:1564–72.

137. Pham A-D, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014;15(1):266.

138. Kontio E, Airola A, Pahikkala T, Lundgren-Laine H, Junttila K, Korvenranta H, et al. Predicting patient acuity from electronic patient records. J Biomed Inform 2014 Oct;51:35–40.

139. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc 2010 Sep;17(5):568–74.

140. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. J Biomed Inform 2012 Aug;45(4):763–71.

141. Sohn S, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. J Am Med Inform Assoc 2011 Dec;18 Suppl 1(Supplement 1):i144–9.

142. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013 May 1;20(3):413–9.

143. Vilar S, Harpaz R, Santana L, Uriarte E, Fried-

man C. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis. PLoS One 2012;7(7):e41471.

144. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clin Pharmacol Ther 2012 Aug;92(2):228–34.

145. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009 May;16(3):328–37.

146. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. J Am Med Inform Assoc 2013 Dec;20(e2):e243–52.

147. Yadav K, Sarioglu E, Smith M, Choi H-A. Automated outcome classification of emergency department computed tomography imaging reports. Acad Emerg Med 2013 Aug;20(8):848–54.

148. St-Maurice J, Kuo MH. Analyzing primary care data to characterize inappropriate emergency room use. Stud Health Technol Inform 2012;180:990–4.

149. Denny JC. Chapter 13: Mining Electronic Health Records in the Genomics Era. PLoS Comp Biol 2012 Dec 27;8(12):e1002823.

150. Cuggia M, Bayat S, Garcelon N, Sanders L, Rouget F, Coursin A, et al. A full-text information retrieval system for an epidemiological registry. Stud Health Technol Inform 2010;160(Pt 1):491–5.

151. Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, et al. Use of computerized algorithm to identify individuals in need of testing for celiac disease. J Am Med Inform Assoc 2013 Dec;20(e2):e306–10.

152. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc 2014 Mar;21(2):353–62.

153. LePendu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. J Biomed Sem 2012;3 Suppl 1:S5.

154. Patel R, Lloyd T, Jackson R, Ball M, Shetty H, Broadbent M, et al. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. BMJ Open 2015;5(5):e007504.

155. Schuemie MJ, Sen E, 't Jong GW, van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. Pharmacoepidemiol Drug Saf 2012 Jun;21(6):651–8.

156. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform 2013;192:572–5.

157. Craig JB, Culley JM, Tavakoli AS, Svendsen ER. Gleaning data from disaster: a hospital-based data mining method to study all-hazard triage after a chemical disaster. Am J Disaster Med 2013;8(2):97–111.

158. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009 Sep;16(5):624–30.

159. Ji Y, Ying H, Dews P, Mansour A, Tran J, Miller RE, et al. A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. IEEE Trans Inf Technol Biomed 2011 May;15(3):428–37.

160. Aoki N, Demsar J, Zupan B, Mozina M, Pretto EA, Oda J, et al. Predictive model for estimating risk of crush syndrome: a data mining approach. J Trauma 2007 Apr;62(4):940–5.

161. Lee J, Maslove DM. Customization of a Severity of Illness Score Using Local Electronic Medical Record Data. J Intensive Care Med 2017 Jan;32(1):38-47.

162. Roth C, Shivade CP, Foraker RE, Embi PJ. Integrating population- and patient-level data for secondary use of electronic health records to study overweight and obesity. Stud Health Technol Inform 2013;192:1100.

163. Schildcrout JS, Basford MA, Pulley JM, Masys DR, Roden DM, Wang D, et al. An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records. J Biomed Inform 2010 Dec;43(6):914–23.

164. Stifter J, Yao Y, Lodhi MK, Lopez KD, Khokhar A, Wilkie DJ, et al. Nurse Continuity and Hospital-Acquired Pressure Ulcers: A Comparative Analysis Using an Electronic Health Record "Big Data" Set. Nurs Res 2015 Oct;64(5):361–71.

165. Westra BL, Savik K, Oancea C, Choromanski L, Holmes JH, Bliss D. Predicting improvement in urinary and bowel incontinence for home health patients using electronic health record data. J Wound, Ostomy Continence Nurs 2011 Jan;38(1):77–87.

166. Yoon D, Park MY, Choi NK, Park BJ, Kim JH, Park RW. Detection of adverse drug reaction signals using an electronic health records database: Comparison of the Laboratory Extreme Abnormality Ratio (CLEAR) algorithm. Clini Pharmacol Ther 2012 Mar;91(3):467–74.

167. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc 2012;2012:901.

168. Hrovat G, Stiglic G, Kokol P, Ojsteršek M. Contrasting temporal trend discovery for large healthcare databases. Comput Methods Programs Biomed 2014;113(1):251–7.

169. Li D, Simon G, Chute CG, Pathak J. Using association rule mining for phenotype extraction from electronic health records. AMIA Summits Transl Sci Proc 2013;2013:142–6.

170. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Comput Biol Med 2006 Dec;36(12):1351–77.

171. Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed 2011 Nov;15(6):823–30.

172. Glasgow JM, Kaboli PJ. Detecting adverse drug events through data mining. Am J Health Syst Pharm 2010 Feb;67(4):317–20.

173. van den Branden M, Wiratunga N, Burton D, Craw S. Integrating case-based reasoning with an electronic patient record system. Artif Intell Med 2011 Feb;51(2):117–23.

174. Skevofilakas M, Zarkogianni K, Karamanos BG, Nikita KS. A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus. Conf Proc IEEE Eng Med Biol Soc IEEE; 2010;2010:6713–6.

175. Zhao D, Weng C. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. J Biomed Inform 2011 Oct;44(5):859–68.

176. Ahmed MU, Funk P. Mining rare cases in post-operative pain by means of outlier detection. IEEE Computer Society; 2011. p. 35–41.

177. Melamed RD, Khiabanian H, Rabadan R. Data-driven discovery of seasonally linked diseases from an Electronic Health Records system. BMC Bioinformatics 2014;15 Suppl 6(Suppl 6):S3.

178. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. J Biomed Inform 2014 Dec;52:199–211.

179. Chen Y, Nyemba S, Malin B. Detecting Anomalous Insiders in Collaborative Information Systems. IEEE Trans Dependable Secure Comput 2012 May;9(3):332–44.

180. Peissig PL, Santos Costa V, Caldwell MD, Rottscheit C, Berg RL, Mendonca EA, et al. Relational machine learning for electronic health record-driven phenotyping. J Biomed Inform 2014 Dec;52:260–70.

181. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov 2010 May;20(3):361–87.

182. Zhang H, Mehotra S, Liebovitz D, Gunter CA, Malin B. Mining Deviations from Patient Care Pathways via Electronic Medical Record System Audits. ACM Trans Manag Inf Syst 2013 Dec;4(4):1–20.

183. Chen Y, Ghosh J, Bejan CA, Gunter CA, Gupta S, Kho A, et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform 2015 Jun;55:82–93.

184. Gupta S, Hanson C, Gunter CA, Frank M, Liebovitz D, Malin B. Modeling and detecting anomalous topic access. IEEE; 2013. p. 100–105.

185. Schuemie MJ, Coloma PM, Straatman H, Herings RMC, Trifirò G, Matthews JN, et al. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med Care 2012 Oct;50(10):890–7.

186. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010 May;26(9):1205–10.

187. Kuhn T, Basch P, Barr M, Yackel T, Medical Informatics Committee of the American College

of Physicians. Clinical documentation in the 21st century: executive summary of a policy position paper from the American College of Physicians. Ann Intern Med 2015;162(4):301–3.
188. Clynch N, Kellett J. Medical documentation: part of the solution, or part of the problem? A narrative review of the literature on the time spent on and value of medical documentation. Int J Med Inform 2015 Apr;84(4):221–8.
189. Weir CR, Hammond KW, Embi PJ, Efthimiadis EN, Thielke SM, Hedeen AN. An exploration of the impact of computerized patient documentation on clinical collaboration. Int J Med Inform 2011 Aug;80(8):e62–71.
190. Jung M, Hoerbst A, Hackl WO, Kirrane F, Borbolla D, Jaspers MW, et al. Attitude of Physicians Towards Automatic Alerting in Computerized Physician Order Entry Systems. Methods Inf Med 2013;52(2):99–108.
191. Gregg EW, Li Y, Wang J, Burrows NR, Ali MK, Rolka D, et al. Changes in diabetes-related complications in the United States, 1990-2010. N Engl J Med 2014 Apr 17;370(16):1514–23.
192. Wollersheim D, Sari A, Rahayu W. Archetype-based electronic health records: a literature review and evaluation of their applicability to health data interoperability and access. HIM J 2009;38(2):7–17.
193. Santos MR, Bax MP, Kalra D. Dealing with the archetypes development process for a regional EHR system. Appl Clin Inform 2012;3(3):258–75.
194. Christensen B, Ellingsen G. Evaluating Model-Driven Development for large-scale EHRs through the openEHR approach. Int J Med Inform 2016 May;89:43–54.
195. Monsen KA, Finn RS, Fleming TE, Garner EJ, LaValla AJ, Riemer JG. Rigor in electronic health record knowledge representation: lessons learned from a SNOMED CT clinical content encoding exercise. Inform Health Soc Care 2014 Oct;:1–15.
196. Daniel C, Ouagne D, Sadou E, Forsberg K, Gilchrist MM, Zapletal E, et al. Cross border semantic interoperability for clinical research: the EHR4CR semantic resources and services. AMIA Jt Summits Transl Sci Proc 2016;2016:51–59.
197. Teodoro D, Pasche E, Gobeill J, Emonet S, Ruch P, Lovis C. Building a transnational biosurveillance network using semantic web technologies: requirements, design, and preliminary evaluation. J Med Internet Res 2012 May 29;14(3):e73.
198. Weis JM, Levy PC. Copy, paste, and cloned notes in electronic health records: prevalence, benefits, risks, and best practice recommendations. Chest 2014 Mar 1;145(3):632–8.
199. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016 Sep;23(5):909–15.
200. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162–73.
201. Bolens M, Borst F, Scherrer JR. Recurrent infections in children observed throughout electronic medical record. Medinfo 1995;8 Pt 1:331.
202. Lovis C, Séroussi B, Hasman A, Pape-Haugaard L, Saka O, Andersen SK. MIE2014 preface. Stud Health Technol Inform 2014;205:v–vi.
203. Milchak JL, Shanahan RL, Kerzee JA. Implementation of a peer review process to improve documentation consistency of care process indicators in the EMR in a primary care setting. J Manag Care Pharm 2012 Jan;18(1):46–53.
204. Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. Expert Opin Drug Saf 2011 Nov;10(6):871–882.
205. Eppenga WL, Derijks HJ, Conemans JMH, Hermens WAJJ, Wensing M, De Smet PAGM. Comparison of a basic and an advanced pharmacotherapy-related clinical decision support system in a hospital care setting in the Netherlands. J Am Med Inform Assoc 2012 Jan;19(1):66–71.
206. Suresh S. Big Data and Predictive Analytics: Applications in the Care of Children. Pediatr Clin North Am 2016 Apr;63(2):357–66.
207. Mohsen MO, Aziz HA. The Blue Button Project: Engaging Patients in Healthcare by a Click of a Button. Perspect Health Inf Manag 2015;12(Spring):1d.
208. Centers for Medicare Medicaid Services CMS. Glossary of eCQI Terms. Available from: https://ecqi.healthit.gov/content/glossary-ecqi-terms
209. Centers for Medicare Medicaid Services CMS. CMS Quality Measures Inventory. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/CMS-Measures-Inventory.html
210. AHRQ. National Quality Measures Clearinghouse (NQMC). Available from: https://www.qualitymeasures.ahrq.gov/index.aspx
211. NQF. National Quality Forum (NQF). Available from: http://www.qualityforum.org
212. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. J Am Med Inform Assoc 2011 Dec 16;18(Suppl 1):i150–i156.
213. Gawron AJ, Thompson WK, Keswani RN, Rasmussen LV, Kho AN. Anatomic and advanced adenoma detection rates as quality metrics determined via natural language processing. Am J Gastroenterol 2014 Dec;109(12):1844–9.
214. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015 Sep;82(3):512–9.
215. Meystre SM, Kim J, Garvin J. Comparing Methods for Left Ventricular Ejection Fraction Clinical Information Extraction. AMIA Summits Transl Sci Proc, CRI; 2012. p. 138.
216. Kim Y, Garvin JH, Heavirland J, Meystre SM. Improving heart failure information extraction by domain adaptation. Medinfo 2013;:185–9.
217. Meystre SM, Kim Y, Heavirland J, Williams J, Bray BE, Garvin JH. Heart Failure Medications Detection and Prescription Status Classification in Clinical Narrative Documents. Stud Health Technol Inform 2015 Aug 21;216:609–13.
218. Meystre S, Meystre SM, Kim Y, Redd A, Garvin JH. Congestive Heart Failure Information Extraction Framework (CHIEF) Evaluation. AMIA Annu Symp Proc 2014:86.
219. Institute of Medicine (US) Roundtable on Evidence-Based Medicine. Washington (DC): National Academies Press; 2009.
220. Budrionis A, Gustav Bellika J. The Learning Healthcare System: Where are we now? A systematic review. J Biomed Inform 2016 Sep 28;64:87–92.
221. Institute of Medicine (US) and National Academy of Engineering (US) Roundtable on Value & Science-Driven Health Care. Washington (DC): National Academies Press (US); 2011.
222. IOM Roundtable on Value & Science-Driven Care and Institute of Medicine. Washington, DC: National Academies Press (US); 2015.
223. Patients Charting the Course: Citizen Engagement and the Learning Health System: Workshop Summary. Washington, DC: National Academies Press (US); 2011.
224. Kuchinke W, Karakoyun T, Ohmann C, Arvanitis TN, Taweel A, Delaney BC, et al. Extension of the primary care research object model (PCROM) as clinical research information model (CRIM) for the “learning healthcare system”. BMC Med Inform Decis Mak 2014 Dec 18;14(1):118.
225. Ethier J-F, Curcin V, Barton A, McGilchrist MM, Bastiaens H, Andreasson A, et al. Clinical data integration model. Core interoperability ontology for research using primary care data. Methods Inf Med 2015;54(1):16–23.
226. Lim Choi Keung SN. Transform: Implementing a Learning Healthcare System in Europe through Embedding Clinical Research into Clinical Practice. Hawaii International Conference on System Sciences HICSS 2015. p. 3176–85.
227. Abernethy AP, Ahmad A, Zafar SY, Wheeler JL, Reese JB, Lyerly HK. Electronic patient-reported data capture as a foundation of rapid learning cancer care. Med Care 2010 Jun;48(6 Suppl):S32–8.
228. Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, et al. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. J Am Med Inform Assoc 2014 Jul;21(4):621–626.
229. Deeny SR, Steventon A. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. BMJ Qual Saf 2015 Aug;24(8):505–15.
230. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Affairs Project HOPE - The People-to-People Health Foundation, Inc; 2008 May;27(3):759–69.
231. Vuokko R, Mäkelä-Bengs P, Hyppönen H, Doupi P. Secondary use of structured patient data: interim results of a systematic review. Stud Health Technol Inform 2015;210:291–5.

Correspondence to:
Stéphane M. Meystre, MD, PhD, FACMI
Medical University of South Carolina
Biomedical Informatics Center
135 Canon St, 4th floor
Charleston, SC 29425, USA
Tel.: +1 843-792-0015
E-mail: meystre@musc.edu