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

S. M. Meystre, C. Lovis, T. Bürkle, G. Tognolli, A. Budrioni, C. U. Lehmann

Medical University of South Carolina, Charleston, SC, USA
Division of Medical Information Sciences, University Hospitals of Geneva, Switzerland
University of Applied Sciences, Bern, Switzerland
Institute of Electronics, Computer and Telecommunication Engineering, Italian Natl. Research Council IEIIT-CNR, Milan, Italy
Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
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, 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 potential 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

I 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 Coorevits 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>
<thead>
<tr>
<th>Principles</th>
<th>Definitions</th>
<th>Impact on Reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle of Information-Privacy and Disposition</td>
<td>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 must importantly use of the data.</td>
<td>• Reasonable protection against any disclosure of patient data. • Patient right to have data expunged or modified.</td>
</tr>
<tr>
<td>Principle of Openness</td>
<td>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.</td>
<td>• Required notification of patients (and raising of awareness) that their data are collected and stored, transmitted, modified, and reused.</td>
</tr>
<tr>
<td>Principle of Security</td>
<td>Collected data must be protected by all reasonable and appropriate measures against loss, degradation, unauthorized access or destruction, use, manipulation, modification, or transmission.</td>
<td>• 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.</td>
</tr>
<tr>
<td>Principle of the Least Intrusive Alternative</td>
<td>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.</td>
<td>• Required analysis of planned reuse of data to avoid infringement or more than minimal interference.</td>
</tr>
<tr>
<td>Principle of Accountability</td>
<td>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.</td>
<td>• Violations of the above principles require the individuals working with reused data (not the primary data collectors) to disclose such events.</td>
</tr>
</tbody>
</table>

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.
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 state (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. 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 fulfill 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].

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 Clyne 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 & development 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 Ilia Lakovidis “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].

I 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 using 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


