Preprocessing structured clinical data for predictive modeling and decision support

A roadmap to tackle the challenges

José Carlos Ferrão[1,2]; Mónica Duarte Oliveira[2]; Filipe Janela[1]; Henrique M. G. Martins[3]

[1]Siemens Healthcare, Rua Irmãos Siemens 1, 2720–093 Amadora, Portugal;
[2]CEG-IST, Centre for Management Studies of Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049–001 Lisbon, Portugal;
[3]Centre for Research and Creativity in Informatics, Hospital Prof. Doutor Fernando Fonseca, IC-19 Venteira, 2720–276 Amadora, Portugal

Keywords
Data mining, data access, integration and analysis, electronic health records and systems, structured data, clinical decision support

Summary

Background: EHR systems have high potential to improve healthcare delivery and management. Although structured EHR data generates information in machine-readable formats, their use for decision support still poses technical challenges for researchers due to the need to preprocess and convert data into a matrix format. During our research, we observed that clinical informatics literature does not provide guidance for researchers on how to build this matrix while avoiding potential pitfalls.

Objectives: This article aims to provide researchers a roadmap of the main technical challenges of preprocessing structured EHR data and possible strategies to overcome them.

Methods: Along standard data processing stages – extracting database entries, defining features, processing data, assessing feature values and integrating data elements, within an EDPAI framework –, we identified the main challenges faced by researchers and reflect on how to address those challenges based on lessons learned from our research experience and on best practices from related literature. We highlight the main potential sources of error, present strategies to approach those challenges and discuss implications of these strategies.

Results: Following the EDPAI framework, researchers face five key challenges: (1) gathering and integrating data, (2) identifying and handling different feature types, (3) combining features to handle redundancy and granularity, (4) addressing data missingness, and (5) handling multiple feature values. Strategies to address these challenges include: cross-checking identifiers for robust data retrieval and integration; applying clinical knowledge in identifying feature types, in addressing redundancy and granularity, and in accommodating multiple feature values; and investigating missing patterns adequately.

Conclusions: This article contributes to literature by providing a roadmap to inform structured EHR data preprocessing. It may advise researchers on potential pitfalls and implications of methodological decisions in handling structured data, so as to avoid biases and help realize the benefits of the secondary use of EHR data.

Correspondence to:
José Carlos Ferrão
Rua Irmãos Siemens 1,
Ed. 3 Piso 3,
2720–093 Amadora, Portugal
Email address: jose.ferrao@tecnico.ulisboa.pt
Telephone: (+351) 214 178 668
Fax: (+351) 214 178 030

Appl Clin Inform 2016; 7: 1135–1153
http://dx.doi.org/10.4338/ACI-2016-03-SOA-0035
received: March 6, 2016
accepted: October 1, 2016
published: December 7, 2016
Citation: Ferrão JC, Oliveira MD, Janela F, Martins HMG. Preprocessing structured clinical data for predictive modeling and decision support – a roadmap to tackle the challenges. Appl Clin Inform 2016; 7: 1135–1153
http://dx.doi.org/10.4338/ACI-2016-03-SOA-0035
Funding
The authors also acknowledge the support from Fundação para a Ciência e a Tecnologia (grant SFRH/BDE/51605/2011), Siemens Healthcare and the Centre for Management Studies of Instituto Superior Técnico (CEG-IST, University of Lisbon).

1. Background

Electronic health records (EHR) have been recognized as a driver for healthcare modernization and widely implemented in multiple settings, producing numerous clinical data repositories with great potential for improving health care delivery and management, and the patient experience [1–3]. Predictive modeling, data mining and clinical decision support represent a central application of the secondary use of EHR data [4], providing proactive insights beyond the scope of human reasoning [2], both from clinical [5] and managerial (e.g. length of stay prediction [6, 7]) perspectives. Despite the potential value of reusing EHR data, there is limited evidence of its impact to date [8], partly due to social and technical barriers [9]. On the one hand, social barriers are mainly related to ethical and regulatory concerns [10], which highly influence the access to data. On the other hand, technical obstacles refer to the retrieval and manipulation of data for the purpose of building decision support tools. While social barriers have received attention from the research community, detailed technical issues remain largely unaddressed. This may be partly explained by significant difficulties in tackling the hurdles associated with the unique nature of health data. These hurdles demand increasing awareness about their specificities and require proper handling so as to avoid arbitrary methodological decisions [11].

In practical terms, many technical challenges arise due to the need to represent data in a matrix format (i.e. "flattened table" [12]), where instances (or data points) are expressed according to a set of features that characterize instances in given context [13], as required for predictive modeling. In this article the term feature is used to designate a (potentially relevant) characteristic of each instance (e.g. a patient episode) in the dataset, and for simplicity purposes features should be regarded as variables typically used in clinical studies, and thus the terms are used interchangeably. Since raw structured EHR data are not natively stored in such format, but rather as database entries based on controlled formats [14], it is necessary to perform multiple retrieval, preprocessing and integration tasks which entail complex methodological options. In order to produce a data matrix format, it is necessary to define what each instance represents and also construct a feature set (defining the lines and columns of the matrix, respectively). While an instance represents a data point, defining a feature set is a complex process which must account for availability, specificity and scope of EHR data, requiring features to be meaningful and measurable for all instances. In clinical settings, constructing a feature set and determining feature values entail a set of challenges that constitute the subject of this article.

The main motivation for this article arose when preparing to carry out these preprocessing tasks in a database of structured EHR data, for which we only found literature with brief general guidelines on preparing and transforming data into a matrix format [15]; and when attempting to build the data matrix from raw data, we realized that we could be leaving core methodological steps omitted or implicit, as well as making arbitrary assumptions and incurring in potential biases and errors. As such, the aim of this article is to contribute to the clinical informatics literature by providing specific guidelines mapped to the main EHR data elements (e.g. diagnoses, procedures and medication, amongst other), providing a roadmap to inform these preprocessing tasks and the transformation of data into a matrix format.

Accordingly, combining our research experience and lessons learned with best practices reported in the literature, this article analyzes the steps, methodological decisions and challenges of handling structured EHR data for predictive modeling, which is herein regarded as a specific application of data mining (which in turn is mostly inspired in machine learning and statistics techniques) consisting in any task of developing models to capture relationships between clinical data and dependent variables and to be able to predict the value of dependent variables from the values of independent variables [16]. It also presents possible strategies to tackle them, so as to support data analysts and researchers upon reusing clinical data. The approaches and methods used to structure clinical data in these systems (i.e., how EHR user interfaces are built to allow users to perform structured data entry) are beyond the scope of this article, since these are typically addressed upstream in the design and implementation of these systems in real-world settings, upon defining user and system requirements. Nevertheless, data preprocessing shall not be dissociated from the techniques used to structure clinical data, and we aimed to capture this relationship by explicitly considering the format of database records resulting from standard data recording mechanisms in preprocessing EHR data.

We also address this relationship in this article by discussing key issues that may impact EHR system designers, implementers and users, whose actions directly impact the structure and content of routinely collected data.

The contents of this article are: section 2 structures the problem at hands and presents a data preprocessing framework to transform EHR data into a data matrix format; section 3 presents the main challenges of using structured EHR data along with strategies to overcome them; section 4 reflects on the data preprocessing framework and compiled challenges, and draws key conclusions.

2. Using structured EHR data for predictive modeling and decision support

2.1 EHR data formats and frameworks

The heterogeneity in data produced across different EHR systems stems from the existence of a wide range of native EHR data formats, which constitute the starting point for a predictive modeling approach based on EHR data. These native formats are expected to comply with certain specifications and share some features, and there have been efforts towards data uniformization: ISO 20514 defines EHR systems as a collection of digital records of patient data [17], ISO 13808 specifies the requirements for EHR systems being "faithful to the needs of healthcare delivery" [18], ISO 21090 define the data types and semantics for representing healthcare concepts [19], and ISO 13606 specifies architecture so as to ensure interoperability [20, 21], for which HL7 has also been a major pillar [22]. In spite of these efforts towards uniformization, there is still significant variability in structure, contents and scope in which these data are stored, which is tightly related to the differences amongst commercial EHR developers and vendors. Besides the ISO regulations, reference models such as the OpenEHR standard have been developed to serve as a groundwork for developing the building blocks of EHR systems [23].

In addition to this difficulty, it is also necessary to account for the fact that the technology used to implement databases of EHR systems also varies: while relational databases are the most frequent methodology [24], alternative technologies have been developed [25] (namely NoSQL [26] and XML databases [27, 28]) as an attempt to accommodate the specificities of health data.

Yet, the characteristics of raw EHR data produced at each provider site highly differ, which may critically hamper the generalization of a data preprocessing framework. Such disparity in health data produced at different sites has been acknowledged, and several initiatives have been put forward with the purpose of enabling systematic reuse of EHR data. For instance, the i2b2 project has aimed to provide a "hive" of building blocks composed by software packages [29] with tools to integrate EHR and genomics data [30]. Similarly, the SHARPn project aims to overcome interoperability and standardization barriers with open-source packages for data migration, structuring and normalization [31, 32]. The EHR4CR project aimed to standardize data collection to promote interoperability and communication for clinical research [33, 34]. Locally-developed data warehouses, such as the one at Vanderbilt University [35], have also appeared. Other relevant initiatives include caGRID [36] and OpenFurther [37]. Overall, these frameworks aim to render EHR data available in standardized formats that enable large scale analysis and integration of data from different sources. However, the use of these frameworks is often precluded by data quality issues and by difficulties in interlinking information models of EHR system and secondary use frameworks. In this and other contexts, reuse of EHR data is only possible by extracting raw EHR data into plain files and then preprocessing these into a format suitable for analysis. It is such preprocessing, specifically, that requires researchers to possess the necessary domain knowledge, employ careful consideration in data manipulation choices, and be aware of the implications of these choices in the analysis.

In such contexts, raw EHR data are extracted from databases and stored into plain text files (e.g. comma-separated values). Since researchers are usually not allowed to directly query EHR databases (to safeguard database performance and protection of commercial data models from vendors), data extraction is usually performed by a technical expert and overseen by review and ethical boards, providing researchers with plain and anonymized EHR data files. These files are not cleaned and

contain structured field values both user-introduced and originating from automated (such as laboratory) systems, by using controlled formats or storing (numerical) results directly in structured fields, respectively. Structure-wise, these files are composed of tuples that are closely linked with the mechanisms for recording (structured) information into the EHR system – pick lists (catalogs), checkboxes, dropdowns buttons and numerical fields – which we describe later on (▶Table 1). These tuples typically contain the instance identifier, the field label, field value, date/time stamp and possibly additional relevant information such as units of measurement. Having all EHR data stored in such tuple-based files, researchers then employ data preprocessing operations to render data usable for predictive analysis. The challenges arising in these data preprocessing tasks are the focus of this article, for which we aim to provide a roadmap and inform researchers on possible strategies to address them, since we did not find literature addressing these challenges.

2.2 Data preprocessing framework – EDPAI

Departing from a raw structured EHR dataset (in plain text files) consisting of a collection of tuples, data preprocessing needs to be performed in several stages which entail multiple challenges. These stages consist in: extract database entries, define features, process data, assess feature values and integrate data elements (standing for EDPAI – ▶Figure 1). The application of this EDPAI framework results in a populated EHR data matrix paired with a label matrix. Similarly to the data matrix, the lines in the label matrix correspond to dataset instances, but its columns represent, in turn, each outcome of interest (i.e. dependent variables). In multi-class and multi-label problems, columns may contain all possible categories, and in multiple applications where only one dependent (outcome) variable is analyzed (which is the case of most predictive modeling problems), the label matrix is represented by a column vector, and in this case it is not necessary to employ an approach to address multi-label problem (possible approaches are described elsewhere [38]). These preprocessing stages are analogous to the rationale of extract-transform-load (ETL) procedures in business intelligence architectures [39], and the EDPAI framework should be regarded as a particular type of ETL process, with inherent specificities and challenges related with the clinical domain and in line with using data from original EHR sources for predictive modelling in the clinical context. Briefly, each of these stages consists upon:

- *Extract database entries*: identifying all necessary EHR data elements (a source of data with a given scope) and querying databases to retrieve all entries of interest, typically using instance identifiers; extraction produces a set of tables, for which the subsequent EDPAI stages are performed separately, integrating these (sub-) matrices at the last stage;
- *Define features*: through a systematic approach, identifying all clinical concepts contained in EHR data and defining features conveying each concept, including its type (numerical or categorical) and mechanism to determine feature value; features must account for different data recording mechanisms specified in section 2.3;
- *Process data*: manipulating the feature set to improve homogeneity and avoid data dispersion by mitigating redundancy (concepts represented with different designations) and granularity (clinical concept is expressed with different levels of detail), which are tackled by combining different features referring to same clinical concept into a single feature;
- *Assess feature values*: determining the value of each clinical feature (variable) for each dataset instance, by querying the extracted database entries according to the feature types and recording mechanisms, as described in section 2.3;
- *Integrate data elements*: concatenating matrices produced from each EHR data element by matching lines of each instance using identifiers, thereby merging matrices side to side; it includes matching each instance with the corresponding line in the label matrix (lines representing instances and columns representing categorical or numerical label values).

In the EDPAI framework, and in general predictive modeling pipelines [40], the construction of a feature set from raw EHR data defines the vector space in which to represent instances and occurs after extracting data referring to a given patient cohort. Without loss of generality, we address this crucial step in the next section, based on the different data recording mechanisms which are pervasive in most typical structured EHR systems.

2.3 Defining features from structured data

Clinical data include multiple data types [41], and different recording mechanisms (▶Table 1 for the main types) are used to capture structured EHR data. Pick lists are used for selections amongst large sets of items (problem lists are a key example [42, 43]). Checkboxes allow the selection of none, one or more items, while dropdowns and radio buttons are used for selecting exactly one item. Numerical fields store values, though records become less controlled in cases of manual input.

Since each recording mechanism produces database tuples differently, it is necessary to adopt a procedure to create features systematically for each mechanism and match them with a feature type (or assuming missingness, as we address later on). Features may be categorical (nominal or ordinal) or numerical (continuous or discrete), which determines how their information can be interpreted (for additional detail, see [44]). The approach to build a feature set and the mechanism to determine feature values from structured data are described in ▶Table 2, linked to the recording mechanisms in ▶Table 1.

▶Table 2 depicts empirical and data-driven procedures to systematically construct features from structured data, being dependent on the conceptual models underlying the data framework and reflecting its scope, granularity and relationships between features. As such, after this first empirical process, researchers must establish the information model and terminologies to use for each EHR data element, particularly by mapping the extracted concepts into the desired terminology using standard schemes such as the UMLS metathesaurus, SNOMED-CT, ICD or CPT for diagnoses and procedures [45], RxNorm for medication [46, 47] or LOINC for laboratory results [48]. In some cases, these reference terminologies can be natively incorporated in the EHR, and researchers should evaluate if the scope and granularity levels meet the requirements of the problem at hands, or if aggregation/simplification is necessary to mitigate data sparsity (e.g., in some studies the high granularity of ICD diagnoses might be excessive).

In order to streamline the construction of a matrix from structured EHR data, ▶Figure 2 presents a procedure to streamline construction of the feature set – and hence of the data matrix – by conveying the systematic feature construction procedures set in ▶Table 2, as a means to reduce manual workload. For pick lists, it is necessary to firstly identify all unique items, flag redundant items and merge these into a single feature, which can be supported by synonyms and hierarchies underlying the terminologies. For entries associated with numerical fields, checkboxes, radio buttons and dropdowns, the process can be facilitated by creating a single source (master) file listing all features defined from EHR fields and specifying parameters of feature name, fields to query, feature type, symbolic encoding (if applicable), how to handle multiple values and whether or not to assume missingness. This file should be used by software with capability of querying database entries and creating a data matrix.

By carrying out the feature construction process, several challenges may arise and preclude the effective reuse of structured EHR data if not properly handled. These challenges stem from the unique nature of health data, which requires proper domain knowledge about which clinical concepts underlie each EHR data element, in order to preprocess it accordingly and – ideally – be able to congregate the different preprocessing tasks into one comprehensive and nearly-automated procedure. It is such dichotomy of specificities of health data and desire to streamline the process that give rise to multiple challenges, which are detailed in the next section, together with strategies to avoid them.

3. Identifying and tackling the challenges

In this section we present the lessons learned from the challenges faced in our experience with a real-world structured EHR dataset (covering approximately 5000 inpatient episodes from medical inpatient wards, over the course of 6 months, from a public hospital with approximately 800 beds), referring to the main methodological decisions along the EDPAI framework, in which we foresaw a potential source of error. In summary, these challenges refer to (1) data gathering and integration, (2) handling feature types properly, (3) combining features, (4) dealing with data missingness, and

(5) handling multiple feature values. ▶Figure 3 links these challenges with EDPAI stages in which they are most likely to occur.

3.1 Data gathering and integration

3.1.1 Challenge 1: Interoperability barriers and instance identification

Difficulties often arise in the process of gathering data of interest (e.g. how to identify which instances are of interest for clinical prediction and decision support models) and upon integrating data from different sources, particularly when different identifiers are used across systems [49]. This is a pervasive issue in health information management [50–52] and may render the dataset incomplete or corrupt. The challenge lies in properly retrieving and integrating all data of interest using a robust match between data elements [50], which is often hampered by the use of different identifiers amongst databases, the use of multiple standards for organizing and transmitting health data in "siloed" applications [34], and the intervention of human agents between systems, which increases the likelihood of errors.

3.1.2 Strategy 1: Robust data retrieval and integration

Extracting and gathering EHR data is typically best achieved with an instance-centered approach, which should follow the trend towards patient-centered EHR systems [53]. A first approach to identify data of interest consists in retrieving unique patient identifiers from different systems by using criteria such as department, dates of admission/discharge and type of episodes, amongst other. Then, integrating different blocks of data can be done by cross-matching patient identifiers (if the same are used) or mapping identifiers and, additionally, cross-matching parameters related to dates of birth, department and other criteria in order to mitigate chances of mismatch. This was the approach followed in our research, and required cross-mapping identifiers used in different systems (e.g. using a different identifier for the laboratory system). In case interoperability between systems is not ensured (and therefore human agents intervene), or when dates/times are not perfectly aligned between systems (time stamps not recorded simultaneously), matching records for data integration will produce numerous mismatches and one should not blindly use all common elements to match instance records. We argue that data retrieval and integration should follow an instance-centered approach, cross-checking fields in a stepwise method: firstly identifying perfect matches and then looking into inconsistencies in order to assess if it is possible to identify and correct the source of error. In our research we found typing errors (related to patient/episode identifiers) to be the main source of mismatch, which required manual inspection. Secondly, numerous mismatches of admission and discharge dates were due to the use of different references amongst systems: the EHR system assumed the admission time in the creation of the patient visit episode, while the admission-discharge-transfer system assumed the date of the first contact of the patient with the hospital (usually with administrative staff, prior to the creation of the episode in the EHR). This manual check of inconsistencies lowers the risk of mismatch, allows for the inclusion of instances that would otherwise be discarded, and identifies causes of mismatch.

3.2 Handling different feature types

3.2.1 Challenge 2: Understanding feature types

EHR contents are composed of different types of information and result in different features types (categorical and numerical [54]) and ranges. Challenges often arise in ascertaining feature types, especially differentiating nominal and ordinal features. The Glasgow coma scale is an example of scale associated with a potential error, in that numerical values are assigned to different categories, yet these only entail ordinal information and should be dealt with accordingly. Furthermore, the choice of prediction models (and also feature selection methods [55–57]) must account for the feature types in order to ensure that a categorical feature (for example, a motor response level of the Glasgow scale, which conventionally represents a numerical score) is not interpreted as a numerical feature and that arbitrary mathematical operations are not performed with such categorical values. For this purpose, researchers must possess full knowledge of statistical procedures for each feature type

and of the operations inherent to each predictive modeling approach, so as to ensure full alignment between the two subjects..

3.2.2 Strategy 2: Matching feature types with model requirements

The identification of feature types should be based on the underlying clinical concept and recording mechanism, by observing if it includes ordinal and cardinal (distance) information. It is particularly important to acknowledge the limitations of unspecified designations ("unknown", "other", "not otherwise specified" or "not elsewhere classified") which despite being problematic are still a commonplace in clinical records [58]. For numerical features, it is fundamental to properly identify units of measure and ensure that necessary conversions are applied. For each feature type, it is important to be aware of the corresponding statistical procedures, for which we present a summary and examples from the clinical domain (▶Table 3).

Considering these statistical procedures, it is essential to understand the way different feature types are handled by each prediction model and which type(s) of features are admitted, of which we provide key examples: neural networks handle discrete features but internally treat them as continuous [59]; regression, support vector machines, neural networks and k-nearest neighbors typically interpret values as numerical; and Bayesian models and decision trees [60] typically accept categorical and numerical features. It is also fundamental to properly choose between classification and regression models according to the categorical (e.g. predicting a diagnosis, a treatment option or a response to a therapy, or other types of classifications) or numerical (e.g. predicting medication dosage, length of stay, costs, or other continuous variables of interest) nature of dependent variables, respectively. The different outputs produced by models (probability estimation, direct class assignment or numerical estimation) should be used properly in each problem.

In predictive modeling studies, feature selection plays a crucial role in mitigating high dimensionality due to the vast number of features (especially binary features resulting from catalogs) arising from structured EHR data. Saeys et al. (2007) [61] provide a comprehensive review of feature selection methods which can be largely applied to EHR data. These methods are divided into filters (independent of prediction models), wrappers (select feature subsets based on resulting model performance) and embedded (feature selection is part of the model-building process). Filter methods require much less computational power and, despite not ensuring optimal results, can be preferable for highly-dimensional clinical data. Within filter methods, the choice of a particular method needs to consider the nature of both the features (independent) and the dependent (outcome) variable(s) at hands. Since several filter methods are based on statistical tests (e.g. chi-squared) and on measures of correlation, these can usually be used for both numerical and categorical features; other methods are based on information theory and require discretization of numerical features, which usually involves testing a number of binarization thresholds. In order to account for the nature of dependent variables, we refer to the review article by Lazar et al. (2012) [62] for a comprehensive view of filter techniques and these should be used for different types of dependent variables.

In spite of restrictions on which type of features models can handle, there are strategies to circumvent these constraints. On the one hand, it is typically necessary to binarize nominal and ordinal features using dummy variables, so as to remove arbitrary ordering and distance measures [61]. On the other hand, discretization converts numerical into ordinal or nominal features, either based on sample distributions (e.g. equal width or frequency binning [62]) or incorporating clinical knowledge to produce clinically meaningful categories, namely by using a clinically relevant discretization threshold, e.g. defining a critical glycaemia level to flag diabetes, a haemoglobin threshold to flag anemia, or a creatinine threshold as an indicator of renal function.

3.3 Combining features

3.3.1 Challenge 3: Feature redundancy and granularity

While EHR systems tend to be increasingly comprehensive and tailored for different care settings [65, 66], there are multiple common concepts that can be expressed with different designations (e.g. heterogeneity in system catalogs [42, 50]) and produce redundancy. A common source of redundancy is the use of different catalogs for diagnoses, medication and procedures, allowing health professionals to use them arbitrarily. Additionally, heterogeneous levels of granularity arise when health

professionals use different levels of detail due to differences in recording practices and in detail needed in each situation [58]. These situations are particularly common for diagnoses, for which the same data may be entered with different designations, and are also observed following the entering of data with different levels of granularity. E.g., when referring to the same clinical status one physician might state that a patient has diabetes mellitus (selecting a more generic diagnosis code), while another might use a more granular designation – diabetes mellitus type II, without complications. When the feature set is being built by identifying unique items, different designations are considered as separate features despite being closely related. Both redundancy and granularity have a negative impact on data sparsity and on predictive power, and should therefore be avoided.

3.3.2 Strategy 3: Towards data uniformization

Domain (clinical) knowledge is crucial to mitigate issues of redundancy and granularity. It is necessary to identify redundant fields and define a mechanism to combine features referring to the same concept into one aggregated feature. When using multiple catalogs, mappings or cross-walks are required for this purpose (e.g. mapping diagnosis catalogs). In fact, catalogs loaded into the system for use in pick-lists play a crucial role since they concretize the terminologies used in the system. Key examples of terminologies consist in the UMLS metathesaurus, the International Classification of Diseases (ICD) and SNOMED-CT, amongst others, and ideally these reference terminologies should be the standard of choice for EHR data entry. However, we did observe in our dataset that health professionals often offer significant resistance to highly controlled and highly granular terminologies, and the creation of (*ad-hoc*) simplified catalogs was found to be the first solution to surpass this adoption hurdle, although with negative consequences for data reuse.

On the one hand, for non-catalog-based recording mechanisms, one may mitigate redundancy by querying redundant EHR fields upon determining feature values (this instruction can be passed to the master file in ▶Figure 2). On the other hand, tackling granularity can be achieved either by aggregating features to the lowest common granular level (causing loss of information) or, alternatively, adding general designations – for instance using clinical knowledge and information from other EHR sources, avoiding arbitrary assumptions.

We recommend careful analysis of the feature set in order to avoid incorporating unnecessary redundancy and sparsity, while accounting for the level of detail required for each predictive modeling problem. This analysis should balance the workload of improving the dataset (namely adding detail to features using clinical knowledge) with the eventual information loss resulting from feature aggregation.

3.4 Data missingness

3.4.1 Challenge 4: The assumption of missingness

Missingness is a transversal problem in data analysis and may have negative consequences on model development [67], potentially leading to biases and information loss [68,69], more so when using multiple EHR data sources, with data being produced by different agents at different frequencies. The first issue related with missing data is whether or not to assume missingness from the absence of a record, since it may have different meanings: it may actually represent a missing entry or the absence of a concept (and thus feature value zero) [70]. Does the absence of a diagnosis or a medication represent the same as the absence of a physiological measurement in terms of data missingness? The second issue consists in deciding how to deal with effectively missing values, either by deleting features and/or instances with missing values or imputing values for missing entries [71].

3.4.2 Strategy 4: Tackling data missingness

The decision of whether the absence of a record implies its missingness requires knowledge of the clinical concept and how information is recorded, depending on both the recording mechanism (from ▶Table 1) and if the field is compulsory (in this case, health professionals are bound to explicitly state if a condition is absent or present). Typically, absence of measurements (e.g. glycaemia, blood pressure and weight) from labeled fields implies, in theory, that feature values are missing. Conversely, the absence of pick-list entries (such as absence of a diagnosis or medication prescription) implies the absence of a concept (and thus a feature value of zero).

For effectively missing values, understanding randomness of missingness patterns is usually a sound first step [72], then deciding how to handle missing data based on this pattern [73]. Deletion or imputation using standard or complex classification methods [74] are widely used techniques. Being difficult to determine the best approach beforehand, several approaches should be explored according to the availability of information, prevalence of missing values and impact on model performance.

3.5 Multiple feature values

3.5.1 Challenge 5: The existence of multiple feature values

Since EHR data represent the evolution of a patient's condition throughout the continuum of care, the same clinical variables are often repeatedly assessed and thereby different label-value pairs are produced for the same field. This situation is highly common for frequent measurements such as vital signs and laboratory exams, where the same clinical variable is assessed multiple times during patient stay. In order to ensure that data matrices only have one value for each feature, it is necessary to decide how to collapse these multiple values into a single value, or set of values, that will convey the desired characteristic while minimizing information loss. However, this process entails restrictions related to feature types, as well as implications of possible mechanisms of collapsing feature values that are sometimes not evident.

3.5.2 Strategy 5: Methods to accommodate multiple feature values

Tackling this challenge is not straightforward and requires careful domain knowledge in order to make informed methodological decisions. As general approaches, one may choose to collapse multiple values into one single value (e.g. calculating the mean or median of multiple blood pressure measurements, or the mode of pain scale levels), define one feature for each measurement (e.g. defining features for first, second and third measurements of the same numerical variable), or a combination of both (e.g. defining features for maximum and minimum values recorded for the same instance, for which extremely high or low values may flag specific conditions, in line with the possible approach for defining binarization thresholds in section 3.2.2). As a means to congregate the different options and associated consequences, ▶ Table 4 aims to inform researchers on possible methods for handling multiple values, along with their advantages and risks.

The decisions of how to accommodate multiple values have great influence on predictive modeling results, since these will determine which information will be captured or discarded from the data matrix. It is typically not suitable to adopt a transversal approach across all features since their values and encoding are intrinsically different, especially when it concerns values that can potentially signal clinical conditions. Additionally, it is also relevant to bear in mind the variability in clinical practice and interpretation of clinical findings, which could be mitigated with the incorporation of rules to detect inconsistencies, signaling them for researchers to ascertain them before proceeding to model building stages.

4. Discussion and conclusions

This article contributes to literature by conveying a roadmap with the main challenges of preprocessing structured EHR data for predictive modeling and decision support, which up to our knowledge have not been systematically addressed in clinical informatics literature. This roadmap is based upon the data preprocessing stages of a generic data preprocessing framework (EDPAI), and was built with lessons learned from our experience and best practices from the literature, proposing possible strategies to mitigate obstacles and pitfalls. The article bridges the gap between general guidelines on the clinical informatics literature [15] and the actual hands-on research work of preprocessing structured EHR data. While the characteristics of EHR systems and the format of raw EHR data produced upon querying EHR databases may be heterogeneous amongst different healthcare providers, the challenges outlined in this article should, in principle, be generally applicable to multiple contexts, particularly the data recording mechanisms and data elements referred throughout this article are pervasive in most settings. We acknowledge that our research and lessons learned

have been based on one particular EHR system, and as such the preparation of structured data should be adapted and mapped to the information models and terminologies underlying EHR systems from other vendors, or locally developed systems. In particular, adaptation amongst different settings may be related with (1) data extraction and integration being dependent on system infrastructure and existing instance (i.e. patient/episode) identifiers, (2) feature construction and value assessment depending on the recording mechanisms (most usually falling within the predefined formats presented in ▶Table 1), and (3) data processing (particularly addressing redundancy and granularity) depending on the catalogs, terminologies and embedded data quality rules.

For scalability, we also presented an empirical data-driven procedure to streamline data preprocessing in creating a data matrix. In this procedure, effort is placed on the initial phase of defining all features from EHR fields, for which changes in the EHR system may be later incorporated with incremental adaptation. To mitigate excessive dependency on each specific EHR implementation, feature set construction should preferably be based on reference terminologies, either by implementing structured data entry based on these terminologies, or by specifying terminologies for which to map EHR records.

Although this article intends to present the main lessons learned from our research experience, it is important to acknowledge that the reuse of EHR data entails such a high degree of complexity that numerous additional challenges may arise from research experience. As such, we believe that the roadmap and best practices presented in this article could be significantly enriched (and modified) by extending the research experience to EHR systems from other hospitals, other vendors and also locally developed systems. Similarly, since the proposed roadmap stemmed from a straightforward mechanism for EHR data reuse (extracting raw plain-text files for subsequent preprocessing and predictive analysis), it may also be significantly improved through the integration of frameworks for data standardization and harmonization (which will likely give rise to additional challenges) and by establishing a collaboration mechanism with an ETL system and architecture to systematically potentiate the reuse of data from a given system. We believe that a more in-depth study of the actual effects of different possible strategies to address challenges may provide insights on the best courses of action to improve the results of predictive analyses.

Two central and closely linked issues arise from this article, which directly influence clinical data reuse: EHR design and configuration, and proficiency of EHR users. Firstly, EHR design requires extensive (health) domain knowledge and proper workflow modeling in order to be successful [75, 76], and directly determines which actions users can perform on the system. Similarly, system implementation determines how faithfully EHR data reproduce the evolution of the patient's health status and treatment processes, and how findings are recorded and revoked. These aspects will have direct impact in data redundancy, granularity, specification of mandatory fields and the existence of multiple feature values, which are directly linked with challenges 3, 4 and 5 of this article. Secondly, training of health professionals in using EHR systems is crucial to determine the format and quality of routinely collected data, coupled with in embedded validation mechanisms [77]. As such, we call attention to the importance of creating awareness of the implications of EHR design and use to the reuse of clinical data, namely by providing proper training and feedback, ultimately aiming for continuous improvement [78].

Due to the aforementioned implications of system design and use for secondary EHR data applications, the target audience of this article is not limited to researchers, but also extended to those involved in system implementation, training and use in routine practice. These stakeholders must regard system usability as a top priority in system design, in order to mitigate the difficulties of structured data entry. For this purpose, we summarize in ▶Table 5 key features for usability and acceptability of structured EHR data entry.

Considering the strong trends in clinical informatics towards structured data entry, we believe that the preprocessing tasks addressed in this article increasingly play a major role in enabling and leveraging the secondary use of EHR data. Therefore, proper consideration of challenges and pitfalls will likely assume an increasingly greater importance, demanding for guidance on how to effectively make sound use of these data. In effect, these challenges may significantly compromise the use of structured data and preclude the realization of the potential value and benefits of EHR systems, if not properly addressed.

Clinical Relevance Statement

The use of structured EHR data for predictive modeling and decision support entails complex preprocessing tasks which are vulnerable to biases and artifacts if implications of methodological decisions are not properly taken into account. To address this matter, this article presents a roadmap of the main challenges of preprocessing EHR data and proposes strategies and best practices to tackle these challenges. The proposed roadmap aims to guide researchers in using structured EHR data for decision support and practical research applications.

Conflicts of Interest

The authors state that they have no conflicts of interest.

Human Subjects Protection

The research underlying the experience and lessons learned presented in this article was conducted with fully anonymized and retrospective EHR data. Data were extracted under the control and supervision of the informatics director and review board of the clinical partner institution, and was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.

Acknowledgments

The authors wish to thank colleagues from Hospital Professor Doutor Fernando Fonseca for close collaboration and availability throughout this research. The authors thank the thorough and insightful comments from anonymous referees on an earlier version of this paper. The authors remain responsible for any omissions and inaccuracies in this article.

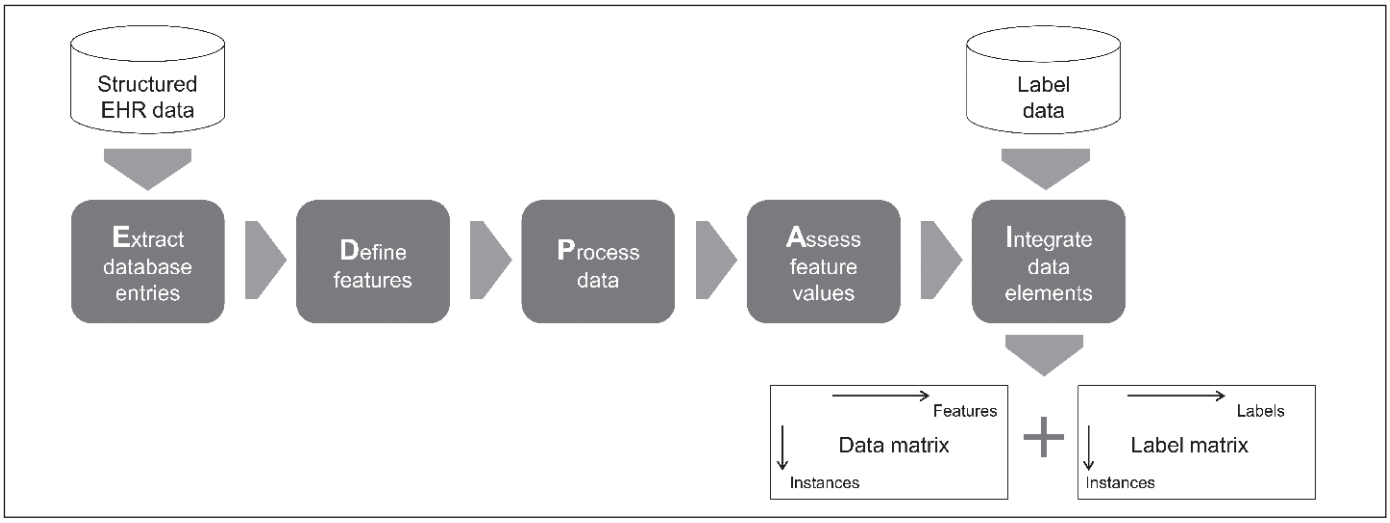

Fig. 1 Building blocks of a predictive modeling framework based on structured EHR data. The gray blocks represent the EDPAI framework. White cylinders and rectangles represent input raw data and resulting matrices, respectively.

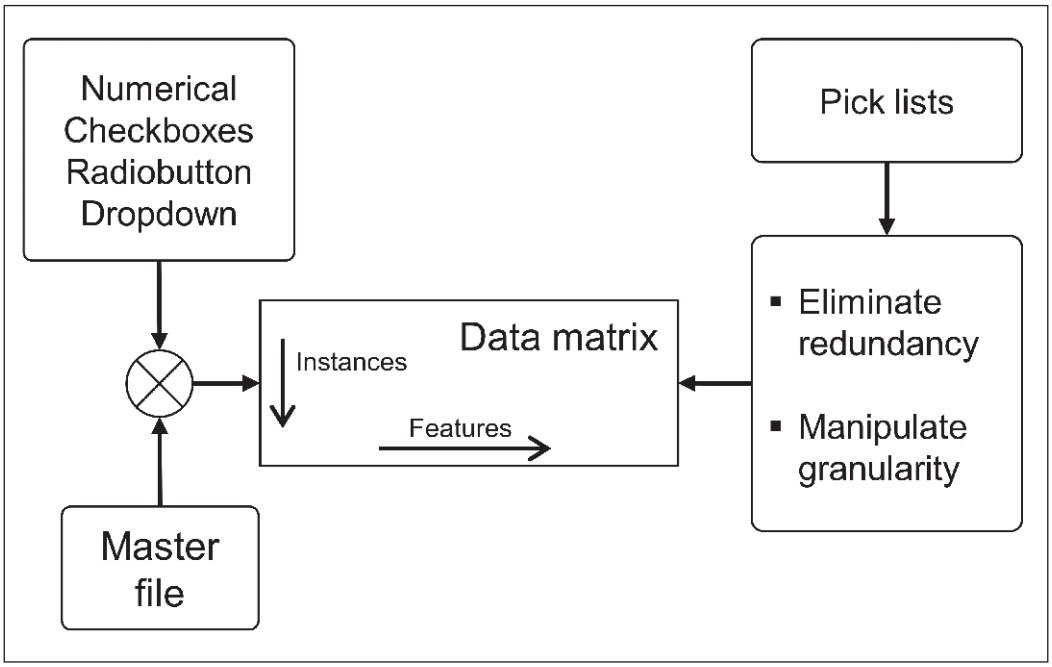

Fig. 2 Procedure to streamline structured EHR data preprocessing to build a data matrix representation.

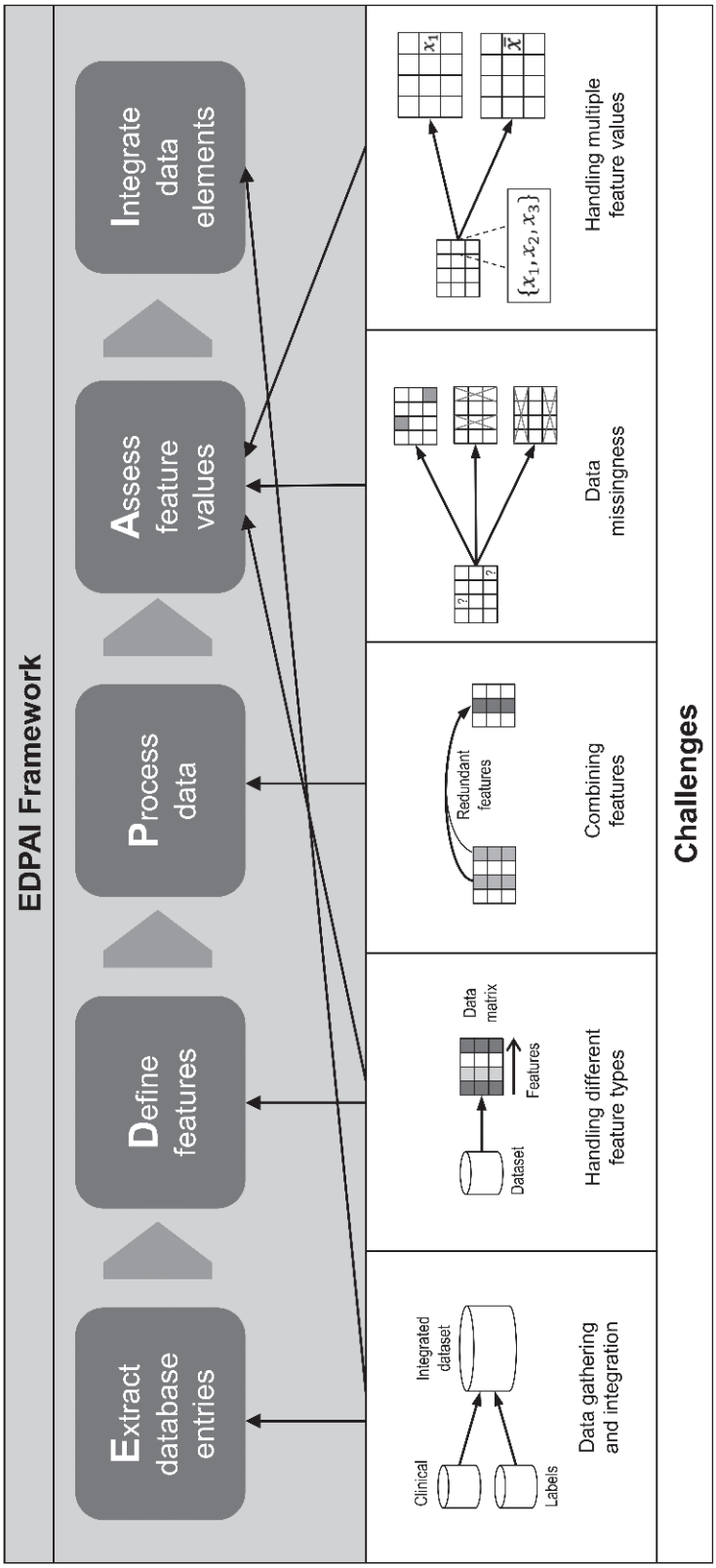

Fig. 3 Schematic representation of the main challenges associated with the use of structured EHR data for predictive modeling. The connections with the EDPAI framework represent the stages where each challenge is most likely to occur.

Table 1 Main types of data recording mechanisms for structured data entry and resulting tuples.

Recording mechanism	Description	Example of tuples
Numerical field	Numerical value stored in a labeled field; tuples may include multiple values with a separator, and additional information for units of measurement	(PatientID; Blood Pressure; 124/67; 2009–02–03 14:55) (PatientID; Glycaemia; 78; mg/dL; 2015–05–21 07:22)
Pick list (catalog)	Selection of one or more items from EHR-embedded catalogs (typically includes a search box)	Diagnoses: (PatientID; Rhabdomyolysis; 2011–01–11 09:18) Drugs: (PatientID; Captopril 50 mg Oral; 2010–03–29 22:07)
Checkbox	Selection of options from a fixed set of items, in a variable number ranging from none to all; field values are usually recorded in tuples with separators	(PatientID; Orientation; Space/Time/Person; 2011–03–29 11:33)
Radio button Dropdown list	Selection of exactly one item from a fixed set of (mutually exclusive) options	(PatientID; Glasgow eye response; 3 (to speech); 2012–11–11 04:09) (PatientID; Catheter type; Central; 2010–07–19 09:25)

Table 2 Procedures to build a feature set from structured EHR data and mechanisms to determine feature values, according to the underlying recording mechanism.

Recording mechanism	Feature definition approach	Mechanism for assessing feature value
Numerical field	Directly define a feature for each field (e.g. define a feature for glycaemia); for multiple values stored in the same field using a delimiter (e.g. blood pressures), define a separate feature for each value	Directly extract field values, parsing values in case of multiple features contained in the same field; use data validation mechanisms to mitigate errors resulting from manual input
Pick list (catalogs)	Define a binary feature for each unique item found in each catalog (e.g. define a binary feature for each relevant anemia diagnoses or each possible medication)	Search for entries of each item for each instance and define value 1 when at least one entry is present in the database, or 0 otherwise.
Checkbox	Define a binary feature for each checkbox option (e.g. define a feature for each possible personal history condition)	Search for entries of corresponding labels; if they exist, define value 1 for features corresponding to selected options, and 0 otherwise.
Radio button Dropdown list	Define a feature for each option; carefully specify the feature type according to the underlying clinical concept (e.g. define a feature for each possible color/appearance of urine samples)	Search for entries of the corresponding label-value pairs, defining feature values using a symbolic encoding scheme

Table 3 Main characteristics and statistical procedures for the different feature types.

Feature type	Explanation	Statistical procedures	Examples
Nominal	Feature values representing labels without ordering or numerical meaning	Mode, entropy, contingency correlation, χ^2	Drain liquid color, diagnosis presence/absence, diet type
Ordinal	Feature values represent labels with ordering but no distance information	Percentiles, rank correlations, t-test, *F*-test	Glasgow scale parameters, functional dependence levels, risk levels
Discrete	Numerical feature values with ordering and distance information, only assuming discrete values in admissible ranges	Mean, standard deviation, Pearson's correlation	Number of labors, blood cell counts, heart and respiratory rates
Continuous	Numerical feature values with ordering and distance information, assuming any value in admissible ranges		Age, blood pressure, glycaemia, oxygen levels

Table 4 Examples of methods to handle multiple feature values and associated risks.

Method	Implementation	Applicable features types	Advantages	Risks
Mean/median or mode	Define one feature using the mean or mode of multiple entries as feature value	Numerical (for mean/median); ordinal, nominal (for mode)	Avoids excessive dimensionality and sparsity	Insensitive to critically high or low values indicating certain conditions or prescriptions
Highest and/or lowest value(s)	Define one feature for highest or lowest value observed within an instance, or define two features to capture both extremes	Numerical, ordinal	Captures critical feature values	Insensitive to the order by which values occurred
First and/or last value(s)	Define one feature for first or last value observed within an instance, or define two features to capture both occurrences	Numerical, ordinal, nominal	Captures values at critical moments of the instance timeline	Insensitive to critically extreme values indicating certain conditions or prescriptions
Improved or deteriorated health status	Define a feature to assess if the patient improved the health status within an instance	Numerical, ordinal	Captures patient evolution along the instance timeline	Insensitive to high or low values above/below certain thresholds indicating certain conditions or prescriptions
Use all observed values	Define a binary (dummy) feature for each possible feature value and define the value 1 for each label within an instance	Ordinal, nominal	Captures all values occurring within an instance	Insensitive to the order of occurrence

References

1. Hripcsak G, Bloomrosen M, FlatelyBrennan P, Chute CG, Cimino J, Detmer DE, et al. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting., J Am Med Inform Assoc 2014; 21(2): 204–211. doi:10.1136/amiajnl-2013–002117.
2. Schneeweiss S. Learning from Big Health Care Data. N Engl J Med 2014; 370: 2161–2163. doi:10.1056/NEJMp1401111.
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; 14(1): 1–9. doi:10.1197/jamia.M2273.
4. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 2010; 48 (Suppl. 6): S106-S113. doi:10.1097/MLR.0b013e3181de9e17.
5. Berner ES. Clinical Decision Support Systems. 2nd ed. New York: Springer; 2007.
6. Rowan M, Ryan T, Hegarty F, O'Hare N. The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artif Intell Med 2007; 40(3): 211–221. doi:10.1016/j.artmed.2007.04.005.
7. Carter EM, Potts HWW. Predicting length of stay from an electronic patient record system: a primary total knee replacement example. BMC Med Inform Decis Mak 2014; 14(26). doi:10.1186/1472–6947–14–26.
8. Chaudry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic Review: Impact of Health Information Technology on Quality, Efficiency, and Costs of Medical Care. Ann Intern Med 2006; 144(10): 742–752.
9. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Inform Assoc 2007; 14(2): 141–145. doi:10.1197/jamia.M2334.
10. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009; 48(1): 38–44.
11. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002; 26(1–2): 1–24.
12. Lin JH, Haug PJ. Data preparation framework for preprocessing clinical data in data mining. Proceedinfs of AMIA Annu Symp; 2006 Nov 11–15; Washington DC, USA. 2006. p. 489–93.
13. Kotsiantis SB, Supervised Machine Learning : A Review of Classification Techniques. Informatica 2007; 31: 249–268.
14. McDonald CJ. Computer-Stored Medical Records: Their Future Role in Medical Practice, J Am Med Assoc 1988; 259(23): 3433–3440. doi:10.1001/jama.1988.03720230043028.
15. Iavindrasana J, Cohen G, Depeursinge A, Müller H, Meyer R, Geissbuhler A, Clinical data mining: a review., Yearb Med Inform 2009; 48 (Suppl. 1): 1–13.
16. Hand DJ, Mannila H, Smyth P. Principles of Data Mining. 3rd edition. Cambridge, USA: MIT Press; 2001.
17. International Organization For Standardization. ISO/TR 20514 Electronic health record – Definition, scope and context. 2005. doi:ISO/TR 20514:2005(E).
18. International Organization For Standardization. ISO 18308 – Health informatics – Requirements for an electronic health record architecture. 2011.
19. International Organization For Standardization. ISO 21090 – Health informatics – Harmonized data types for information interchange. 2011.
20. International Organization For Standardization. ISO/EN 13606 – Health Informatics – Electronic Health Record Communication. 2010.
21. Santos MR, Bax MP, Kalra D. Building a logical EHR architecture based on ISO 13606 standard and semantic web technologies. Stud Health Technol Inform 2010; 160(Pt 1): 161–165.
22. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006; 13(1): 30–39. doi:10.1197/jamia.M1888.
23. Beale T, Heard S. OpenEHR Architecture Overview. 2006.
24. Atzeni P, De Antonellis V. Relational database theory. Redwood City, USA: Benjamin-Cummings Publishing; 1993.
25. Lee KK, Tang WC, Choi KS. Alternatives to relational database: comparison of NoSQL and XML approaches for clinical data storage. Comput Methods Programs Biomed 2013; 110(1): 99–109. doi:10.1016/j.cmpb.2012.10.018.
26. Cattell R. Scalable SQL and NoSQL data stores. ACM SIGMOD Rec 2011; 39(4): 12–27. doi:10.1145/1978915.1978919.
27. Stalidis G, Prentza A, Vlachos IN, Maglavera S, Koutsouris D. Medical support system for continuation of care based on XML web technology. Int J Med Inform 2001; 64(2–3): 385–400. doi:10.1016/S1386–5056(01)00195–2.
28. Catley C, Frize M, A prototype XML-based implementation of an integrated "intelligent" neonatal intensive care unit. Proceedings of the 4th Int IEEE EMBS Spec Top Conf Inf Technol Appl Biomed; Apr 24–26 2003; Birmingham, UK. 2003. p. 322–325. doi:10.1109/ITAB.2003.1222543.

29. Gainer V, Hackett K, Mendis M, Kuttan R, Pan W, Phillips LC, et al. Using the i2b2 hive for clinical discovery: an example. Proceedings of AMIA Annu Symp; 2007 Nov 10–14; Chicago, USA. 2007. p. 959.
30. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010; 17(2): 124–130. doi:10.1136/jamia.2009.000893.
31. 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; 45(4): 763–771. doi:10.1016/j.jbi.2012.01.009.
32. Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, et al. The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. Proceedings of AMIA Annu Symp; 2011 Oct 22–26; Washington DC, USA. 2011. p. 248–56.
33. 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; 53: 162–173. doi:10.1016/j.jbi.2014.10.006.
34. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, et al. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011; 44 (Suppl. 1): S94-S102. doi:10.1016/j.jbi.2011.07.007.
35. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014; 52: 28–35. doi:10.1016/j.jbi.2014.02.003.
36. Oster S, Langella S, Hastings S, Ervin S, Madduri R, Phillips J, et al. caGrid 1.0: an enterprise Grid infrastructure for biomedical research. J Am Med Inform Assoc 2008; 15(2): 138–149. doi:10.1197/jamia.M2522.
37. Bradshaw RL, Matney S, Livne OE, et al. Architecture of a federated query engine for heterogeneous resources. Proceedings of AMIA Annu Symp; 2009 Nov 14–18; San Francisco, USA. 2009. p. 70–4.
38. Tsoumakas G, Katakis I, Vlahavas I. Mining Multi-label Data. In: Mainon O, Rokach L, editors. Data Mining and Knowledge Discovery Handbook. New York: Springer; 2010. p. 667–685.
39. Wu L, Barash G, Bartolini C. A Service-oriented Architecture for Business Intelligence. Proceedings of the IEEE Int Conf Serv Comput Appl (SOCA); 2007 Jun 19–20; Newport Beach, USA. 2007. p. 279–285. doi:10.1109/SOCA.2007.6.
40. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J Biomed Inform 2014; 48: 160–170. doi:10.1016/j.jbi.2013.12.012.
41. Pietka E. Large-Scale Hospital Information System in clinical practice. Int Congr Ser 2003; 1256: 843–848. doi:10.1016/S0531–5131(03)00458–8.
42. AHIMA Work Group. Problem List Guidance in the EHR. J AHIMA 2008; 82(9): 73–77.
43. Holmes C, The Problem List beyond Meaningful Use Part I: The Problems with problem Lists. J AHIMA 2011; 82: 30–35
44. Moshkovich H. Rule induction in data mining: effect of ordinal scales. Expert Syst Appl 2001; 22(4): 303–311. doi:10.1016/S0957–4174(02)00018–0.
45. Cimino JJ, Review paper: coding systems in health care. Methods Inf Med 1996; 35(4–5): 273–284.
46. Liu S, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005; 7(5): 17–23. doi:10.1109/MITP.2005.122.
47. Bennett CC. Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records. J Biomed Inform 2012; 45(4): 634–641. doi:10.1016/j.jbi.2012.02.011.
48. Huff SM, Rocha RA, McDonald CJ, De Moor GJE, Fiers T, Bidgood WD, et al. Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary. J Am Med Inform Assoc 1998; 5(3): 276–292. doi:10.1136/jamia.1998.0050276.
49. Doan A, Halevy A, Ives Z. Principles of Data Integration. 1st ed. Morgan Kaufmann; 2012.
50. Brazhnik O, Jones JF. Anatomy of data integration. J Biomed Inform 2007; 40(3): 252–269. doi:10.1016/j.jbi.2006.09.001.
51. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. Yearb Med Inform 2008: 91–101.
52. Giuse D. Health information systems challenges: the Heidelberg conference and the future. Int J Med Inform 2003; 69(2–3): 105–114. doi:10.1016/S1386–5056(02)00182-X.
53. Donnelly WJ. Viewpoint: patient-centered medical care requires a patient-centered medical record. Acad Med 2005; 80(1): 33–38.
54. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008; 77(2): 81–97. doi:10.1016/j.ijmedinf.2006.11.006.
55. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res 2003; 3: 1157–1182.

56. Liu H, Motoda H, Setiono R, Zhao Z. Feature Selection: An Ever Evolving Frontier in Data Mining. in: JMLR Work Conf Proc 2010; 10: 4–13.
57. Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997; 1: 131–156. doi:10.1016/S1088–467X(97)00008–5.
58. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998; 37(4–5): 394–403.
59. Lippmann R. An introduction to computing with neural nets. IEEE ASSP Mag 1987; 4(2): 4–22. doi:10.1109/MASSP.1987.1165576.
60. Quinlan JR. Decision trees and decision-making. IEEE Trans Syst Man Cybern 1990; 20(2): 339–346. doi:10.1109/21.52545.
61. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23(19): 2507–2517.
62. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinform 2012; 9(4): 1106–1119. doi:10.1109/TCBB.2012.33.
63. Suits DB. Use of Dummy Variables in Regression Equations. J Am Stat Assoc 1957; 52: 548–551. doi:10.1080/01621459.1957.10501412.
64. Liu H, Hussain F, Tan CL, Dash M. Discretization: An Enabling Technique. Data Min Knowl Discov 2002; 6: 393–423.
65. Chen C, Garrido T, Chock D, Okawa G, Liang L. The Kaiser Permanente Electronic Health Record: transforming and streamlining modalities of care. Health Aff (Millwood) 2009; 28(2): 323–333. doi:10.1377/hlthaff.28.2.323.
66. Mäenpää T, Suominen T, Asikainen P, Maass M, Rostila I. The outcomes of regional healthcare information systems in health care: a review of the research literature. Int J Med Inform 2009; 78(11): 757–771. doi:10.1016/j.ijmedinf.2009.07.001.
67. Heitjan DF. Annotation: what can be done about missing data? Approaches to imputation. Am J Public Health 1997; 87(4): 548–550.
68. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338: b2393. doi:10.1136/bmj.b2393.
69. Gorelick MH. Bias arising from missing data in predictive models. J Clin Epidemiol 2006; 59(10): 1115–1123. doi:10.1016/j.jclinepi.2004.11.029.
70. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013; 1(3): 1035. doi:10.13063/2327–9214.1035.
71. Allison PD. Missing Data. SAGE Publications, Inc.; 2001.
72. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59: 1087–1091. doi:10.1016/j.jclinepi.2006.01.014.
73. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JM, Finkelstein SN. Missing data in medical databases: impute, delete or classify?. Artif Intell Med 2013; 58(1): 63–72. doi:10.1016/j.artmed.2013.01.003.
74. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 2010; 50(2): 105–115. doi:10.1016/j.artmed.2010.05.002.
75. Windle T, McClay JC, Windle JR. The impact of domain knowledge on structured data collection and templated note design. Appl Clin Inform 2013; 4(3): 317–330. doi:10.4338/ACI-2013–02-CR-0008.
76. Rosenbloom ST, Stead WW, Denny JC, Giuse D, Lorenzi NM, Brown SH, et al. Generating Clinical Notes for Electronic Health Record Systems. Appl Clin Inform 2010; 1(3): 232–243. doi:10.4338/ACI-2010–03-RA-0019.
77. Hoerbst A, Ammenwerth E. Electronic Health Records. A Systematic Review on Quality Requirements. Methods Inf Med 2010; 49(4): 320–336. doi:10.3414/ME10–01–0038.
78. Cresswell KM, Bates DW, Sheikh A. Ten key considerations for the successful implementation and adoption of large-scale health information technology. J Am Med Inform Assoc 2013; 20(e1): e9–e13. doi:10.1136/amiajnl-2013–001684.
79. Walji MF, Kalenderian E, Piotrowski M, Tran D, Kookal KK, Tokede O, et al. Are three methods better than one? A comparative assessment of usability evaluation methods in an EHR. Int J Med Inform 2014; 83(5): 361–367. doi:10.1016/j.ijmedinf.2014.01.010.
80. Walji MF, Kalenderian E, Tran D, Kookal KK, Nguyen V, Tokede O, et al. Detection and characterization of usability problems in structured data entry interfaces in dentistry. Int J Med Inform 2013; 82(2): 128–138. doi:10.1016/j.ijmedinf.2012.05.018.