Os artigos de investigação do Centro de Estudos de Gestão do Instituto Superior Técnico (CEG-IST) destinam-se a divulgar os resultados da investigação realizada pelos seus membros.

The working papers of the Centre of Management Studies of IST (CEG-IST) are aimed at making known the results of research undertaken by its members.

Pedidos de informação sobre estes artigos, ou relativos a investigação feita pelo Centro devem ser enviados para:

Enquiries about this series, or concerning research undertaken within the Centre should be sent to:

Presidente do CEG-IST
Instituto Superior Técnico
Av. Rovisco Pais,
1049-001 Lisboa
Portugal
e-mail: cegist@tecnico.ulisboa.pt

Centro de Estudos de Gestão IST

Artigo de Investigação / Working Paper

ISSN 1646-2955

Nº 5/2013

Clinical coding based on structured EHR systems – a supervised learning approach using routinely collected data

José C. Ferrão, Mónica D. Oliveira, Filipe Janela, Henrique M. G. Martins

Abstract

Objectives: clinical coding is an essential process whereby health data is indexed using standard terminologies for billing and reporting. To mitigate workload and errors, partial automation has been seek using free-text data from electronic health records (EHR), but its use is difficult to generalize in contexts with different scopes and languages. In this work, we propose an approach to support clinical coding using fully structured EHR data.

Methods: we propose a methodology encompassing EHR data processing to define a feature set and a supervised learning approach to predict ICD-9-CM code assignment. We employ a fast correlation-based filter to reduce dimensionality and binary relevance method to transform the multi-label problem into multiple binary classification problems. Four supervised learning models – decision trees, naïve Bayes, logistic regression and support vector machines – were tested and compared.

Results: tests performed with a real dataset yielded F1 scores of 0.48, 0.58, 0.54 and 0.53 for decision trees, naïve Bayes, logistic regression and support vector machines, respectively. Performance varies greatly across codes and appears to be related to the clinical concepts underlying each code rather than to the supervised learning method employed.

Conclusions: the use of structured EHR data to support clinical coding shows promising results. The analyses carried out in this work indicate lines for further improvement, namely data quality assessment and improvement and the inclusion of expert knowledge in model revision and validation.

1. Introduction

The translation of health data to a standardized terminology has long been a challenge for healthcare providers. This process, designated as clinical coding, consists in indexing the information regarding clinical conditions and health problems, as well as the medical services provided to a patient during a given episode, making use of a classification scheme [1]. The purpose of clinical coding is twofold: while it initially aimed to create a reference terminology according to which health data should be reported, allowing for statistical and health profiling studies, these classification schemes have subsequently been used as a basis for billing and financing purposes. The responsibility of correct and timely coding falls entirely on healthcare providers, and therefore clinical coding has become a crucial process in the workflow of healthcare providers, especially for its financial implications [2].

Multiple coding schemes are currently in force, amongst which the International Statistical Classification of Diseases and Related Health Problems (ICD) stands out for indexing morbidity and mortality data. Multiple versions of ICD are presently implemented in different health systems. The 9th Revision, Clinical Modification of ICD (ICD-9-CM) stemmed from the ICD-9 version and was locally modified in the United States by refining the original ICD scheme, incorporating more categories and adding a third volume concerning the classification of medical procedures [2]. ICD-9-CM is the standard coding scheme in Portugal, where this study was carried out, as well as in numerous other countries besides the USA. Structurally, ICD-9-CM consists of a hierarchical scheme with five main levels – chapters, sections, categories, subcategories and subclassifications. For diagnosis ICD-9-CM codes, a certain health condition may be represented by a code, consisting in a combination of 3, 4 or 5 alphanumeric characters. According to guidelines, coding professionals should use the highest possible level of detail upon assigning codes to patient episodes – for instance, it is not possible to assign a 3 digit code for which 4 or 5-digit codes are available [3].

In practical settings, coding is carried out by certified professionals (who may be physicians, as happens in Portugal) following patient discharge. To this end, the entire medical record is reviewed to identify information describing reasons causing contact with health services and the patient's clinical condition throughout the episode, as well as medical and surgical procedures provided. This information is looked up in

coding guides in order to produce a set of ICD codes to assign to each episode. Naturally, high patient volumes and complexity of medical record data result in a resource-intensive and error-prone workload for healthcare providers. With increasing pressures for cost containment, efficiency and quality control, healthcare providers have been motivated to obtain significant efficiency and cost reduction gains by streamlining clinical coding, which can be potentiated by the increasing adoption of electronic health record (EHR) systems [4].

EHR systems are considered fundamental tools in pursuing quality, safety and efficiency gains. These systems have been progressively implemented in numerous healthcare settings [5], [6], despite several challenges and resistance in their adoption [7]–[9]. The implementation of EHR systems has greatly changed the paradigm of data capture, not only enhancing data storage, retrieval and sharing [10], but also placing multiple data capture points along care provision [11] and thereby producing massive amounts of data with potential for decision support [12], despite eventual concerns in using EHR data for such purpose [13], [14]. Information created through the systematic use of EHR systems has potential value not only for clinical decision support, but also to assist management. In fact, EHR systems enable symbolic representations of data so that large volumes of information can be automatically processed using controlled terminologies. Considering the shift from paper to electronic formats, there is scope for exploiting automated data processing mechanisms as a means to provide, to a certain extent, support to the coding process. This motivation is also related to the notion that coding is, in a considerable proportion, repetitive and relatively straightforward, and therefore a partial automation of clinical coding can generate organizational gains. Within this context, this area has captured the interest of numerous researchers that have focused their attention mostly in the development of clinical coding support tools based upon free-text and narrative data.

In spite of the wealth of knowledge stored in EHR systems, the use of free-text (narrative) format is a major limitation, hampering automated use of health data for decision support [15]. While human readers are able to read and extract concepts from medical texts, this knowledge is “locked” in narratives and thereby not readily usable [16]. Furthermore, while developments in medical information extraction rely on natural language processing (NLP) methods and have proven successful in numerous contexts, including clinical texts, and numerous authors have proposed methodologies to support clinical coding using free-text data and

NLP methods [17], their use in clinical free-text data has lagged behind in relation to other contexts, most likely due to intrinsic characteristics of these texts such as the lack of sentence structure, the use of short-hand lexical units and frequent misspellings pose significant challenges to information extraction [18]. Despite the myriad of coding support studies found in the literature, there are still multiple barriers to the adoption of these approaches in current practices, mostly arising from issues in generalizing NLP-based methods in the clinical domain [19], [20], especially when clinical texts exhibit poor quality and contain acronyms, ambiguities and uneven structures, and are produced in languages for which NLP lacks off-the-shelf resources. In this study, we investigate the use of structured data from a patient-centered EHR system to develop models to support clinical coding, surpassing most challenges associated with processing text data and pursuing generalizability across healthcare settings. For this purpose, our research was based on a hospital implementation of the EHR system SOARIAN® in Portugal, which started going live in medical wards in early 2012. At the time of development of this study, the EHR implementation had reached a substantial level of maturity, with the far majority of inpatient episodes recorded in the hospital's admission-discharge-transfer system exhibiting EHR system records. This system was used by health professionals on a routine basis, as the main support for recording narrative and structured health data. Clinical documentation is, naturally, produced using European Portuguese.

This paper is structured as follows: section 2 presents key literature addressing clinical coding support based on EHR data. Next, we present a methodology to use structured EHR data to support the coding process, firstly describing in section 3 the EHR systems structure, the definition of the feature set and the mechanism to process data in its EHR-native format, and then in section 4 the supervised learning models (decision trees, naïve Bayes, logistic regression and support vector machines – SVM) used to predict ICD code assignment, as well as methods for feature selection and model regularization. Section 5 presents results from the experiments carried out in this study. Lastly, key aspects of our methodology and the results obtained with different models and directions for future improvements are discussed in Section 6.

2. Related work

While earlier tools aimed to facilitate code lookup, more recent approaches have aimed to actively interpret information contained in medical records to propose a set of codes to be validated by coding professionals [4]. Literature addressing coding support (including not only multiple ICD versions, but also SNOMED) contains highly variable approaches and contexts of application [17], and up to our knowledge materials used to extract medical information have been invariably composed of free-text. The underlying clinical documents have ranged from admission notes [21] to radiology reports [22]–[28], discharge summaries [29]–[39] and entire medical charts [40]–[43], and therefore the scope and detail of information differ. The level of document structuring also varies, as does the quality of narrative texts, since most texts are produced through dictation and commonly may have reasonable semantic and orthographic quality, especially when compared to texts written on-the-spot. Moreover, the vast majority of studies focused on, and used resources dedicated to, the English language, with a minority using French [32], [38], [43], [44] or German [30]. We did not find studies were based on Portuguese-written data. The scope of clinical conditions also varies greatly across literature, ranging from limited sets (respiratory [28], cerebrovascular [35] or coronarography exams [29]) to heterogeneous internal medicine episodes [32], which reflect the range of codes considered, from five [40] or six [35] to twenty [45] or fifty [46] codes, and with one study considering more than 1400 codes [38]. Previous studies have presented a narrow scope by not providing generalizable approaches considering the entire range of ICD codes.

Since all available studies were based on narrative data, numerous NLP tools were employed for text processing. MedLEE [47] is at the core of several studies and other resources such as MetaMap [48], NegEx [49] and UMLS dictionaries [50] have been employed for specific text processing tasks. Despite their extensive use, NLP are: (1) language-dependent and thus are difficult to reproduce amongst languages; and (2) most often require extensive annotated corpora developed specifically for the NLP task at hands, which is itself a resource-intensive process. Therefore, NLP tools are difficult to "adapt, generalize and reuse" [20].

Two major trends are being observed in the way clinical coding is modeled and tackled: (1) a group of studies makes use of NLP tools to extract concepts and directly suggest codes, and (2) a more frequent trend combining NLP tools to extract concepts to achieve a feature-vector representation (typically using bag-of-

words models) and subsequently applying machine learning algorithms to predict code assignment. The type of classifiers include SVM [22], [25], [40], [45], [46], [51], naïve Bayes [38], [42], decision trees [28], ridge regression [40], [46] and k-nearest neighbors [22], [34], [43]. The reported performance values are rather discrepant, although proper comparison is not possible due to the use of different metrics (namely accuracy, precision, recall, F1 scores, area under the receiver-operating characteristic curve and kappa values) and the fact that studies address coding challenges of different complexity, which precludes a fair comparison of results [17]. Notwithstanding comparison issues, some studies provide encouraging results, suggesting the potential use of supervised learning models to develop models for coding support.

In the context of our study, the use of NLP is deemed impractical due to the lack of generalizability of these methods and limited availability of resources for the Portuguese language, as well as the particular characteristics of medical data found in our context, with texts having quality, ambiguity and structure issues. Hence, the motivation for this study is to leverage the existence of novel EHR systems with the bulk of medical information stored in structured formats, with a high level of integration in clinical practice indicating availability of information with relevance for clinical coding. We seek to investigate the use of such natively-structured data with the main purpose of surpassing the challenges of clinical text processing, also making use of supervised learning approaches to take advantage of the continuously populated database of EHR episode data.

3. Methods

In this work, we follow a supervised learning paradigm to support clinical coding. This section describes the two fundamental building blocks of the proposed methodology: the data structuring framework whereby a set of variables (features) is defined to represent EHR data, and the supervised learning approach used to develop models to predict ICD coding. The application of such a methodology requires firstly the definition of a feature set according to which each episode is represented, by developing a data matrix representation in line with typical machine learning frameworks [52], and subsequently the development of prediction models based on these variables.

3.1. Data structuring

The methodology proposed to support clinical coding consists essentially in modeling the coding process to make use of information contained in medical records and predict which codes to assign to each episode. Since the entire medical record is consulted in clinical coding, it is necessary to comprise all relevant information upon operationalizing these variables, henceforth designated as features, based on the characteristics of the EHR system in use. However, data are heterogeneous and stored in different formats [53], and therefore this stage is not straightforward and requires a careful definition of features according to the underlying clinical concepts. In this section, the characteristics and data contents of the EHR system used in this work are described, and modelling options taken to build the feature set are described in section 3.2.

3.1.1. The EHR system

This study makes use of data from a commercial EHR system – SOARIAN® [54] – which is implemented in numerous hospital settings (and also smaller scale providers) throughout the world. Its most relevant feature consists in its patient-centered nature, whereby the majority of clinical data related to a given patient is congregated in the same database. Therefore, coherence between different data elements (e.g. demographic and medication data) is ensured, avoiding data fragmentation amongst different databases with risks of information loss. Data integration is particularly relevant in this work since clinical coding makes use of the entire medical record. SOARIAN® is segmented into several data elements with different scopes, which are depicted in Figure 1. Besides demographic data, the remaining components of the SOARIAN® system may be grouped into two sets: the first group consists in system components primarily meant to characterize clinical conditions, health status and personal history. The second group encompasses information on medical services provided during each episode.

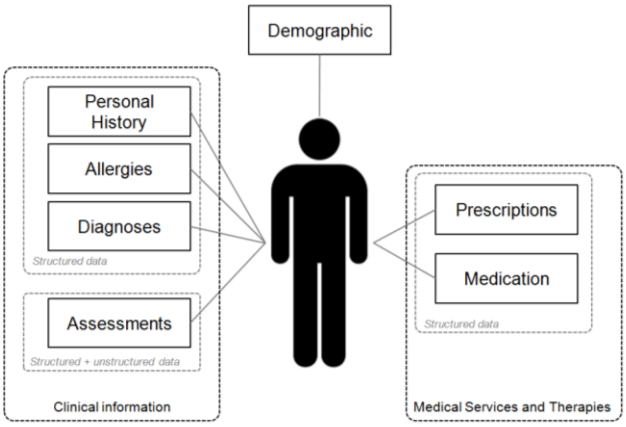

Figure 1 - Patient data components contained in the EHR system – SOARIAN®.

In the first data group, data elements include personal history, identified allergies, laboratory tests results and diagnoses assigned by physicians. This latter topic is specifically relevant and contains all diagnoses found relevant by physicians as motivating patient admission, explaining the patient's health status, and is deduced from objective observation and findings from diagnostic testing. Additionally, SOARIAN® stores detailed clinical information through structured forms, designated by assessments, consisting in sets of labeled fields filled in by health professionals and stored in a given point of the each episode's timeline, which can be consulted or edited. These assessments may be considered, in terms of content, as equivalent to typical narrative documentation. Admission notes and discharge summaries are examples of such assessments. Other assessments contain more specific information, as for instance the case of "Respiration" assessments, storing parameters related to mechanical ventilation and other breathing aids, and the "Vital signs" assessment tailored for recording parameters such as heart rate, blood pressure and pain and coma scales.

All assessments are composed of structured (checkboxes, buttons, dropdown lists or numeric values) and free-text fields to accommodate narrative notes. Upon system implementation, hospital stakeholders define key aspects to be represented in structured formats. Still, structured fields contained in assessments are of different nature (e.g. nominal, ordinal or numerical) and thus need to be properly handled so as to avoid variable misinterpretation and biased models. Moreover, there may be a certain level of redundancy across different assessments, since the same clinical information may be reported in different occasion. All these

issues must be taken into account upon defining the feature set from structured data, and more importantly, in specifying which data fields should be queried to assess feature values in order to fill in the data matrix.

The second group of SOARIAN® data elements comprises care services provided to patients, containing prescriptions of medical and nursing procedures, ranging from diagnostic and laboratory tests to more complex actions such as thoracentesis, as well as medication administered during care provision. All items found in these two subgroups are selected from catalogs embedded in the EHR system and, therefore, this information is fully structured, expressed in controlled vocabularies, though redundancy may still occur in these data elements and the number of unique items may be quite high. In this case, the natural approach is to define a binary feature for each item in the catalog and to perform simplifications to minimize dimensionality and unnecessary details.

Clinical information stored in the EHR system is rather heterogeneous, with data being recorded in different formats, by different professionals and with different frequencies. In order to establish the basis upon which to build prediction models, it is necessary to map relevant information to a feature set characterizing each episode, while accounting for the specificities of the information being manipulated, so as to avoid inconsistencies and incorrect interpretation.

3.1.2. Modeling EHR data

Since the coding process consists in a thorough review of the whole medical record in order to identify all concepts referring to conditions, diagnoses and medical services occurring in a given episode, we define, in the feature set, a binary variable for each possible concept that may be identified in a given episode, assigning 1 if that concept is present and 0 otherwise. This rationale is used in studies based on textual data by using NLP techniques to identify such relevant concepts in medical narratives and evaluating each episode for their presence or absence. In this study, we adopt an analogous strategy for purely structured EHR data elements which are selected from catalogs, i.e., for diagnoses, prescriptions, medication, personal history and allergies. In this sense, we consider a binary feature for each item in these catalogs. Before defining binary variables, simplifications are performed in prescriptions (e.g. removing distinctions of X-ray exams by number of incidences) and medication data is stripped out of dosage and administration method and items representing

combinations of therapies existing elsewhere in the catalog are eliminated. Diagnoses, in turn, are selected using three different redundant catalogs, which were mapped to the same catalog (with validation from clinicians) in order to eliminate redundancy. Lastly, allergies were harmonized to reflect allergen active substances (rather than drug commercial names).

In addition to the above data elements, data recorded through assessments required more complex processing. Firstly, it was necessary to perform an exhaustive listing of all fields and identify clinical concepts conveyed by these fields. In numerous cases, different fields pointed to the same concept and only one variable was considered in such cases, combining information from different fields. While some concepts are directly traduced in model features (e.g. blood pressure and Glasgow coma scale) and assume the value of these fields, other (binary) concepts may be inferred not only from information contained in fields (e.g. presence of urinary catheter: yes/no), but also from the existence of data in other fields (e.g. information on the date of catheter insertion is not a relevant datum *per se*, yet informing on the presence of urinary catheter). Moreover, it was also necessary to consider the different nature of the data fields – numerical (continuous or discrete) or categorical (nominal or ordinal) – and the frequency of occurrence. For categorical variables, dummy variables were created to eliminate artifacts of arbitrary ordering. Conversely, the values of numerical features were directly captured from the values recorded in assessment fields. As for accommodating several values of the same variable occurring in a given episode, dummy variables assumed the value 1 for all feature values occurring in a given episode, while numerical variables were, in turn, split into two variables for the maximum and minimum values occurring in each episode, aiming to minimize information loss.

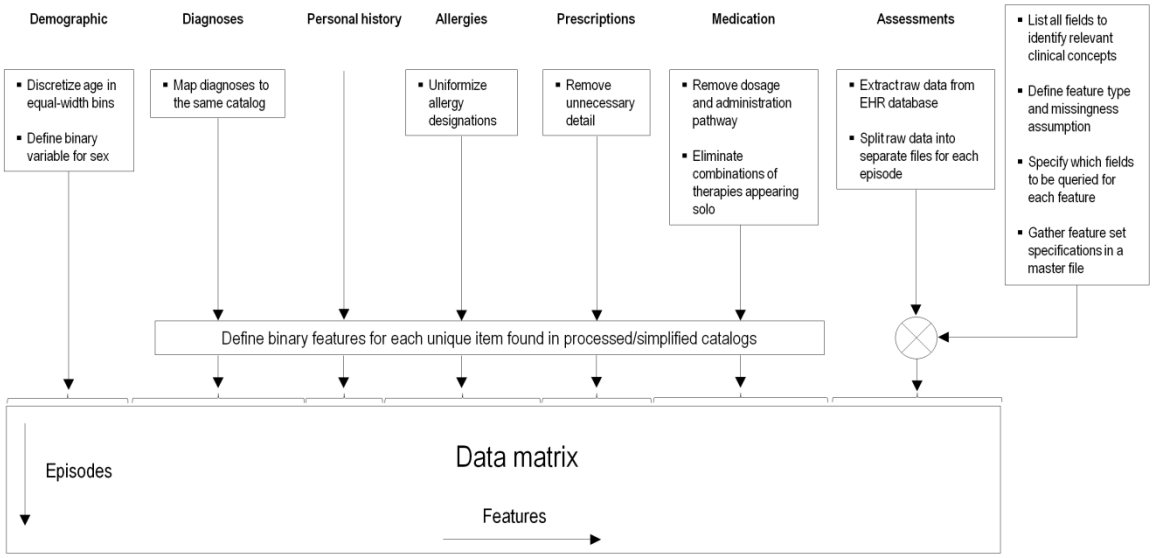

Figure 2 – Data structuring and data matrix construction: mapping raw episode data stored in the EHR database to achieve a data matrix representation.

The process of mapping information directly from the EHR database, where it is stored in its raw format, to a data matrix representation that allows developing prediction models is depicted in Figure 2. For purely binary data (diagnoses, personal history, allergies, prescriptions and medication), this process consists in identifying unique catalog items, simplifying these items in line with the problem scope, and then populating the data matrix according to the presence or absence of items in each episode. For assessment data, the mapping task was fully automated once the feature set is defined through a master file and, in order to maximize method adaptability, that task was performed by a MATLAB® algorithm which reads such file to inspect EHR in its raw format (MySQL) database and populates the corresponding feature values in the data matrix. In such configuration, the process of building a data matrix directly from the database is streamlined and easily adaptable to changes in the dataset or in the structure and contents of the EHR assessments.

3.2. Models for ICD code prediction

In this work, the aim is to support ICD coding by taking advantage of data from past episodes available in the EHR database. These historical data consist not only of clinical data from each episode, but also of ICD

codes assigned to each episode, and therefore it is natural to adopt a supervised learning paradigm for which numerous models are established in the literature [52]. In this paradigm, a given model is adjusted (i.e., trained) using a set of instances and the corresponding outcomes, then being able to perform predictions on new, unseen instances. Having built a dataset of inpatient episodes (where instances are represented according to a feature set, which in this work is defined in line with the previous section), this work addresses key aspects of a supervised learning approach, namely concerning feature selection to reduce dimensionality, the selection of a supervised learning model, parameter adjustment and use of regularization to prevent overfitting, and also the approach by which the multi-label nature of clinical coding is handled. Thereby, methods and techniques from different fields of study were hereby combined so as to account for all the specificities underlying the use of structured EHR data to support clinical coding.

The supervised learning framework includes numerous models, which have the common purpose of modeling (learning) the relationship between feature values and corresponding outcomes from a set of instances (the training set) and then extrapolating these trends for unseen instances. These algorithms highly differ in aspects such as the type of features (numerical/categorical) handled, the rationale for modeling feature-outcome relationships, the training algorithms to fit model parameters to historical data and interpretability of results. In this context, the choice of supervised learning models needs to account for multiple aspects such as (1) that models should be able to accommodate both numerical and categorical data, (2) should be scalable to contexts with high dimensionality and dataset cardinality, and (3) model outputs should preferably allow interpretation, which is essential for validation by coding professionals. As there is no axiomatic guideline as to which model should be applied in each problem, we implemented and tested four models to predict the assignment of ICD codes: decision trees, naïve Bayes classifiers, logistic regression models and SVM. These models have been widely used across the literature, including several studies aiming to support clinical coding with free-text data, and therefore exhibited potential applicability in our context.

In order to handle the multi-label character our coding challenge – wherein each instance may be assigned one or more labels (codes), as opposed to supervised learning problems where each instance is assigned precisely one label. To address this matter, this study employed a problem transformation approach, whereby a multi-label classification problem is decomposed into multiple single-label classification problems [55].

This approach has been used in other studies, namely through a binary relevance method in which the clinical coding problem is decomposed by creating a binary classifier for each code, predicting its inclusion (or not) in the suggested set of codes to assign to each episode. For scalability purposes, methods considering all possible label combinations were found inadequate due to the high number of possible combinations and low representativeness of each combination in training sets.

We hereafter present some key features of each supervised learning algorithm, as well as the methods used to perform feature selection and, for logistic regression, model regularization.

3.2.1. Supervised Learning Models

We hereby briefly describe the four supervised learning models used in this study, employed to perform classification deriving from the binary relevance problem transformation method. These models highly differ in the approach to model the relationship between feature values and outcomes, using measures based on entropy, likelihood or distances, as well as in the rationale for making predictions based on feature values.

Decision trees were the first models considered in this study, which are widely used especially in contexts with a high proportion of categorical features. These models consist in recursively partitioning the dataset based on feature values in the dataset and are typically represented in a tree-like structure wherein each internal node represents a data splitting step [56]. In order to classify an instance, its feature values are evaluated according to the order specified by the decision tree model, then classifying that instance using the label assigned to leaf nodes. Training decision tree models from a dataset consists in determining which attributes should be selected as splitting criteria at each node, in order to minimize generalization errors. Multiple criteria to select splitting attributes have been proposed [57]. Decision trees are also prone to overfitting, caused by excessive adjustment to training data and consequent loss of generalizability. To mitigate this issue, two complementary approaches may be adopted: pre-pruning to avoid excessive tree growth, and post-pruning by discarding tree branches after model-growing has stopped to achieve an optimal tree structure.

While decision trees perform data classification using measures related to entropy and information gain (underlying splitting criteria), an alternative, yet very common, approach to classification consists in

estimating *a posteriori* probabilities $P(C_k|x)$ of an instance belonging to class k given its feature values x. These probabilities may be estimated using generative or discriminative approaches. An example of the former is the naïve Bayes classifier wherein priors $P(C_k)$ and likelihood values $P(x|C_k)$ are firstly estimated in order to compute $P(C_k|x)$ for each class using Bayes rule:

$$P(C_k \mid x) = \frac{P(C_k)P(x \mid C_k)}{P(x)} \tag{1}$$

The naïve Bayes classifier is a well-established supervised learning model in which prior probabilities may be empirically obtained from the proportion of instances belonging to each class in the training set. Conversely, the (otherwise intractable) class-conditional probability estimation is simplified with the assumption of conditional independence. Under this assumption, naïve Bayes classifiers model class-conditional joint probabilities as the product of the class-conditional probabilities of each feature x_j, then using the following decision rule:

$$C_k \leftarrow \arg\max_k P(C_k)\prod_j P(x_j \mid C_k) \tag{2}$$

Applying eq. (2), each new instance is assigned to the class that maximizes the posterior probability. Although features may not be conditional independent, naïve Bayes classifiers tendentially perform well also in contexts where this assumption is violated. This observation is deemed relevant in the context of this work, since several features, for instance related to certain diagnoses and prescriptions, may exhibit statistical correlations naturally arising from their relationship in the clinical context.

Despite having a rationale similar to that of naïve Bayes based on *a posteriori* class probabilities, logistic regression represents, in turn, discriminative models in which these probabilities are estimated directly from training data. These models are tailored especially for problems with dichotomous outcome variable in that the probability for assigning the positive class ($k = 1$) is modeled using a logistic link function, as follows:

$$P(C_{k=1} \mid x) = \frac{1}{1 + e^{-(w_0 + \sum_{j=1}^{N} w_j x_j)}} \tag{3}$$

$$P(C_{k=0} \mid x) = \frac{e^{-(w_0 + \sum_{j=1}^{N} w_j x_j)}}{1 + e^{-(w_0 + \sum_{j=1}^{N} w_j x_j)}} \tag{4}$$

Training logistic regression models consists in estimating the parameters w_0 and w_j which best fit the data, typically using maximum likelihood estimation [58]. Using such coefficients estimates, fitted models are used to predict binary outcomes variables using the following decision rule based on eqs. (3) and (4):

$$\frac{P(C_{k=1} \mid x)}{P(C_{k=0} \mid x)} > 1 \Rightarrow k = 1 \tag{5}$$

In the clinical coding problem, eq. (5) represents the conditions in which ICD-9-CM codes are assigned to a given unseen episode according to the model fitted for each ICD code using a training set. Since the probabilities of a two-class problem must add up to 1, eq. (5) is equivalent to considering a decision threshold of 0.5. This threshold may, however, be adjusted according to compensate for class imbalances.

The fourth algorithm, SVM, adopt a different approach to classifier training. These models aim to define a hyperplane that is able to separate data from different classes and then use such hyperplane as decision boundary to classify instances based on their positions relatively to the boundary. This hyperplane is determined by maximizing the functional margin defined by the nearest training instances. Since training sets are usually not linearly separable, it is frequent to map the feature space to another space with different (possibly infinite) dimension. This mapping is made by applying a kernel function $\phi(\boldsymbol{x})$, which can be linear, polynomial or a radial basis function, amongst other. SVM classifiers (represented as vectors $\boldsymbol{w}$) are obtained from training data as the solution of the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \tag{6}$$

$$\text{s. t.} \qquad y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \tag{7}$$

Parameters ξ and C represent incorrectly classified instances and the penalty (cost) applied to these misclassifications, respectively. The choice of penalty C is reflected in the width of the separating hyperplane

margin, and thus in the amount of misclassified instances (which, in this context, consist in episodes incorrectly assigned a given code or, conversely, episodes not assigned a suitable code). As in other machine learning models, training SVM classifiers involves manipulating parameters to maximize model generalizability while allowing some training instances to be misclassified. In the case of the linear kernels used in this work, only parameter C is subject of manipulation.

Despite using different procedures to model coding data, all four methods implemented in this study yield, for each episode, a set of binary outputs (as many as the number of ICD codes considered) indicating which codes should be assigned to that episode. By comparing these predictions with known outcomes (in test data), model performance is evaluated using the metrics detailed in section 4, which are widely used in studies in the area. Naturally, the number of ICD codes considered in a given context varies according to the range of codes found in the dataset (which, *in extremis*, may amount to over 14.000 codes existing in the ICD-9-CM catalog), yet, more importantly, on the selection of a subset of codes composing the greater bulk of code occurrences, which have considerable representativeness in the dataset. In effect, the imbalance of numerous classes may be a drawback of a supervised learning approach, which may preclude its applicability and seriously hamper its results [59]. In the context of clinical coding, the practical impact of data imbalance is partially alleviated by the fact that, in most cases, the majority of code occurrences are found in a relatively small subset of codes with reasonable occurrence rates and, consequently, a supervised learning approach has potential to support a broad range of the clinical coding activity. A more detailed analysis of this topic is covered later on in this paper.

3.2.2. Feature selection and regularization

The problem addressed in this study poses challenges to supervised learning algorithms arising from high data dimensionality and sparsity, as the approach used in this study to produce a data matrix from raw EHR data results in large feature sets. Moreover, large feature sets are also prone to multicolinearity and other statistical artifacts. To address these challenges, this study makes use of feature selection to reduce the dataset dimensionality, whereby a subset of features is utilized. To this end, two approaches are available: filter methods that select subsets independently from the chosen learning model, and wrapper models that select

subsets based on the resulting model prediction accuracy (and are, therefore, dependent on the chosen algorithm) [60]. Since wrapper methods require fitting prediction models to candidate subsets in order to assess their predictive accuracy, these methods become computationally heavy for high-dimensional datasets. Therefore, in this study we have adopted filter methods for their lower computational requirements.

Filter methods typically aim to determine subsets of relevant features, within which redundancy may also occur. We employed a feature selection method proposed by Yu and Liu (2004) which accounts for both relevance and redundancy – the fast correlation-based filter (FCBF), which handles feature redundancy explicitly and has been shown to perform efficiently on highly dimensional data [61]. FCBF specifies relevant features using a measure of symmetrical uncertainty (SU):

$$SU(X,Y) = 2\left[\frac{IG(X \mid Y)}{H(X) + H(Y)}\right] \tag{8}$$

This measure informs on the correlation between a feature and the independent variable and is based on information gain (IG) and entropy (H) principles. Feature redundancy is, in turn, evaluated according to the concept of Markov blankets. To apply the FCBF method, it is necessary to define a threshold value for SU that determines which features are selected by the method. The selection of this threshold value in the range between 0 and 1 depends on the specificities of each context (higher threshold values will result in smaller feature subsets) and should be adapted accordingly. Within our methodological framework, feature selection methods are applied prior to the model-training stages. Given the multi-label character of coding support, feature selection must be performed for each code (i.e., for each binary classification task), and hence the importance of employing a computationally efficient feature selection method.

Besides issues related to computational load, high-dimensional datasets also carry an additional challenge for logistic regression models, due to their susceptibility to overestimate coefficients β through standard maximum likelihood. To surpass this matter, selection and shrinkage methods are employed to regularize models by penalizing large coefficients. In particular, lasso (least absolute shrinkage and selection operator) is an efficient method to avoid overestimating coefficients while also performing feature selection [62] (since some coefficients may be shrunk to zero, which does not happen with conventional shrinking methods such as

ridge regression). Lasso estimation is performed by adding a penalty factor to the maximum likelihood estimator for *N* instances (without penalizing the independent term β_0):

$$\max\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \quad (9)$$

In eq. (9), λ represents a tunable regularization parameter determining the shrinking magnitude. The typical approach consists in fitting logistic regression models using multiple values of λ and selecting the value exhibiting better classification performance, particularly using cross-validation [63].

4. Results

This section presents the results obtained from an application of the proposed methodology to a dataset obtained from real-world settings. A dataset of inpatient episodes was extracted from the EHR database, referring to patients admitted in the medical departments of Internal Medicine, Pneumology, Nephrology, Infectiology and Gastroenterology of a large public hospital in the area of Lisbon, Portugal, during the first semester of 2013. This study analyzes the potential of the proposed methods to assist the assignment of diagnosis ICD codes based on routinely collected structured data.

4.1. Dataset

The dataset was composed of 5089 inpatient adult episodes. After applying the processing mechanisms described in Figure 2, dimensionality amounted to 5023 features, some of which – particularly those obtained from assessment fields – exhibited missing values, since their specification included assuming that a value was missing when no record regarding that feature was found. In this dataset, 203 (4.04 %) features exhibited missing values, and were discarded to avoid biases, resulting in a final dimensionality of 4820 features.

As expected, the occurrence of ICD codes was found to be highly imbalanced. In our dataset, those 4820 features corresponded to a total of 39273 code occurrences composed of 2272 unique ICD-9-CM diagnosis codes. Dataset imbalance is observed in the proportion of episodes carrying each ICD code (Figure 3) and in the number of ICD-9-CM codes with very few occurrences (e.g. 860 of the 2272 ICD codes in this dataset were assigned only once). Considering the computational load required to make use of data with high

dimensionality and cardinality, most of the analyses conducted in this study focused on the 50 most frequent ICD-9-CM codes, which accounted for nearly 50% of total code occurrences in the dataset.

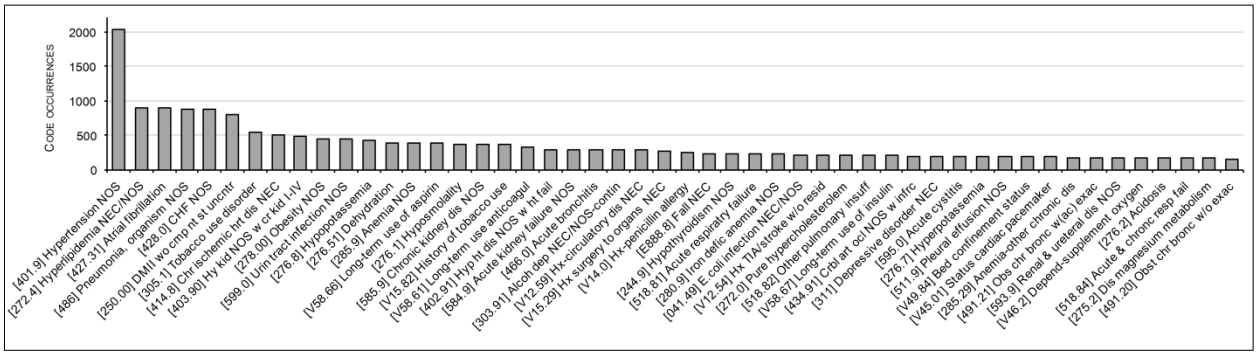

Figure 3 – Number of code occurrences for the 50 most frequent ICD-9-CM codes.

Table 1 – Proportion of code occurrences of the 50 most frequent ICD-9-CM codes, grouped by chapter.

Chapter	Section description	% Top 50
001-139	Infectious And Parasitic Diseases	1,16%
240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	23,33%
280-289	Diseases Of The Blood And Blood-Forming Organs	4,30%
290-319	Mental Disorders	5,59%
390-459	Diseases Of The Circulatory System	28,74%
460-519	Diseases Of The Respiratory System	12,47%
580-629	Diseases Of The Genitourinary System	7,91%
V01-V91	Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services	15,23%
E000-E999	Supplementary Classification Of External Causes Of Injury And Poisoning	1,27%

4.2. Figures of merit

The four supervised learning models proposed in this study were built using 5-fold cross-validation, whereby the dataset is randomly partitioned into 5 non-overlapping subsets, then using 4 of these subsets (the training set) to train each model and testing the models' ability to predict ICD code assignment on the remaining subset (the test set). This mechanism is performed 5 times, using one of the 5 subsets as test set at a time. In this study, training and test sets are composed of 4072 and 1017 instances, respectively. To assess model performance on the test set, measures based on the number of true positives (TP – number of episodes correctly assigned a given code), false positives (FP – number of episodes incorrectly assigned a given code)

and false negatives (FN – number of episodes for which a given code was missed) were employed. These measures of interest are also named precision (P), recall (R) and F1 score (F1), being computed for each ICD code *i*:

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{10}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{11}$$

$$\mathrm{F1}_i = \frac{2P_iR_i}{P_i + R_i} \tag{12}$$

In multi-label classification problems, aggregation of performance measures may be performed using macro or micro-averaging, with the former consisting in averaging measures obtained for each code, whereas the latter consists in first adding TP, FP and FN counts obtained for each code and then use these total values in the formulas presented in eqs. (10-12) [64]. In order to perform a refined analysis of model behavior, performance is analyzed as macro-averaged values, as well as for each code individually.

4.3. Experiments and model performance

In order to analyze key aspects related with the proposed methodology, we firstly assess the capability of supervised learning models to predict ICD code assignment using the original feature set resulting from our data processing approach. Then we evaluate the benefit of introducing the FCBF feature selection method and of diagnosis detail in model performance. Finally, we analyze the behavior of model performance as the number of ICD codes covered increased, i.e. including ICD codes with very low occurrence rates.

4.3.1. Impact of FCBF feature selection, regularization and diagnosis detail

We firstly analyze model performance using the full set of features obtained using the data processing framework proposed in this study, composed of all 4820 features without missing values – hence without employing feature selection or regularization (in the case of the logistic regression) – which constitutes our

base case. Then, we introduce feature selection and model regularization and compare these results with the base case. Upon developing the supervised learning models, it was necessary to adjust parameters specific of each model. Parameter manipulation was performed for each binary classifier (i.e. for each ICD code), as follows:

i. In decision trees, parameter manipulation to minimize overfitting was performed in two stages. Firstly, pre-pruning was employed during the model-building process by imposing a minimum number of data points in each leaf node (considering a minimum of 1, 3 and 5 data points). After each tree model had been fully grown, post-pruning was then performed by progressively trimming trees from the bottom up and selecting the best pruning level based on the F1 score obtained in the test set;

ii. For naïve Bayes, parameter manipulation consisted in varying the classification threshold (for the *a posteriori* probability) from 0 to 1 in steps of 0.05. This manipulation aimed to compensate the fact that most codes were highly imbalanced and, therefore, prior values for positive classes tended to be low, thus lowering the *a posteriori* probability computed using Bayes' rule (eq. 1);

iii. In logistic regression models, parameter manipulation was similar to the one performed in Bayesian classifiers, considering classification thresholds from 0 to 1 in steps of 0.05;

iv. Lastly, for SVM classifiers, a linear kernel was employed since radial basis function kernels exhibited poor results and higher computational effort. Therefore, only the C parameter found in eq. 6 was adjusted, using C values between 10^{-2} e 10^{5}, with unitary exponent increments.

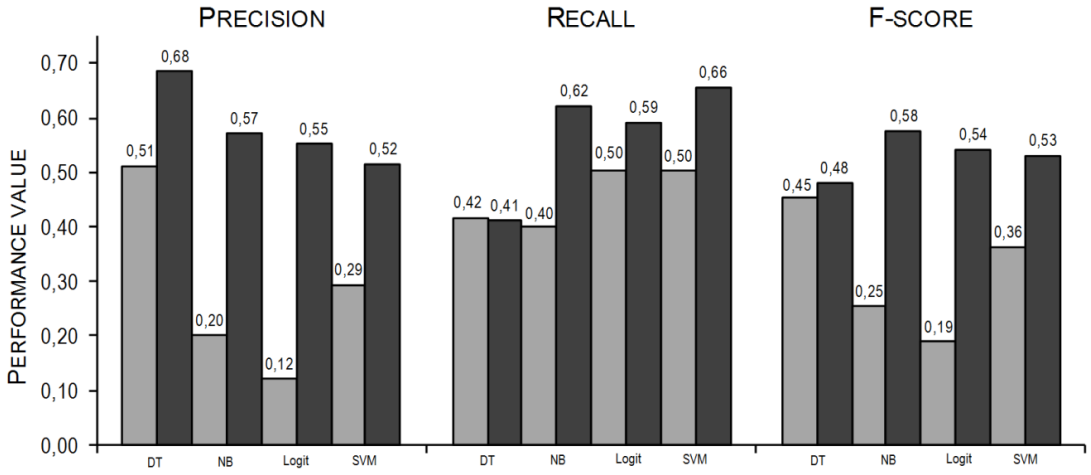

Figure 4 – Macro-averaged precision, recall and F1 scores obtained with decision trees (DT), naïve Bayes (NB), logistic regression (logit) and support vector machine (SVM) models for the 50 most frequent ICD-9-CM codes. Light gray represents results without feature selection and regularization. Dark gray represents results with FCBF feature selection and lasso regularization.

Figure 4 shows macro-averaged performance figures obtained with the 4 supervised learning models – decision trees (DT), naïve Bayes (NB), logistic regression (logit) and support vector machines (SVM), with and without feature selection and regularization. These values were obtained by selecting model configurations (i.e. feature subset and model-specific parameters – minimum number of instances in leaf nodes and post-pruning level for decision trees, classification threshold for naïve Bayes and logistic regression, regularization parameter λ for logistic regression as well, and the soft margin parameter C for SVM) according to the F1 scores, averaged across the 5 folds, yielded by each configuration.

The performance values obtained without feature selection and regularization were relatively low for coding support purposes, which were thought to be closely related with the high dimensionality of our dataset. Decision trees were, as expected, still able to partially overcome this high dimensionality because of its intrinsic feature selection process embedded in model training stages, referring to the selection of splitting nodes (in this study, based on the Gini index criterion), aiming to identify most statistically relevant splits while mitigating model overfitting by restricting the minimum number of instances in leaf nodes. However, a high computational effort was required to train decision tree models using the full feature set, since a large number of variables had to be tested as candidates for splitting criterion at each node. Naïve Bayes classifiers,

in turn, suffer from higher dimensionality mostly upon estimating class-conditional probabilities, especially when using multivariate multinomial distributions in cases where some feature-class combination values do not appear in the training set (which, for higher dimensionality, is much more likely to occur). As for logistic regression models, since these are not tailored to handle datasets with such high number of predictors (due to the small sample biases), model-fitting yields over-estimated factors (with orders of 10^{16}) and fitted values perfectly separate positive from negative examples. In such case, models are fitted to the data and not to the trend, resulting in a heavy loss of generalizability, which is not desirable in a context where the aim is to extrapolate knowledge from past episodes into making predictions for code assignment in future, unseen episodes.

Base case results therefore claimed for the use of feature selection and regularization. FCBF proved to be able to handle high dimensional datasets efficiently although still being able to handle continuous features if discretized [65]. In this study, we made use of the FEAST Toolbox implementation for Matlab® [66], with which a feature subset of relevant and non-redundant features was selected for each code. To specify the SU threshold with which to select features, we did a series of feature selection experiments using SU threshold values ranging from 10^{-1} to 10^{-4} with decreasing exponential steps of 0.25. We observed that, for all 50 most frequent codes, the number of selected features stabilized before reaching the lower threshold value and, therefore, no further features were considered relevant (and non-redundant) by FCBF. These largest feature subsets selected with the FCBF method were then taken for the subsequent model training stage. Figure 5 exhibits the number of features selected for the 50 most frequent ICD codes, showing that dimensionality was greatly reduced, considering the original set of 4820 features.

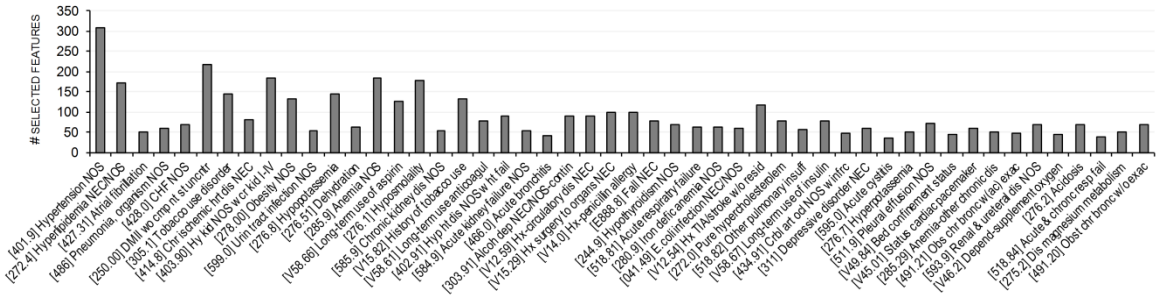

Figure 5 – Number of features selected for each of the 50 most frequent ICD codes using the FCBF algorithm.

We adopted a stepwise approach of adding one variable to the models at a time, according to the feature order (decreasing order of relevance) found in FCBF output. Hence, we trained each of the 4 supervised learning methods for each code starting with the most relevant feature and adding one variable at a time, then training models and assessing performance for each of created model. Because of the relatively long model training times and the number of parameter optimizations inherent to each model trained, we truncated the maximum number of features at 50 for codes having more than 50 features selected by FCBF.

In addition to introducing FCBF, we also added regularization in logistic regression using lasso estimation (eq. 9). This technique not only allows regularizing estimated factors, but also introduces additional feature selection by shrinking some coefficients to zero when higher penalties are used. The magnitude of regularization is controlled by the parameter λ in eq. 9, of which we used 100 λ values in a geometric sequence, the largest one determined as the minimum value of λ yielding all coefficients equal to zero and with a smallest-to-largest λ ratio of 10^{-4}.

By introducing FCBF and regularization, model performance significantly improved for nearly all models, except for a non-significant loss in recall for decision trees. These results provide evidence on how crucial feature selection and regularization methods become in contexts of high dimensionality, which in this case is particularly relevant considering the use of structured EHR data. Naturally, decision trees benefited less with FCBF than the other three methods, as these do not include an intrinsic feature selection approach. Still, the

former showed improvement in terms of overall performance and naturally required much lower computational effort by searching fewer candidate features at each splitting node.

As for overall performance in terms of F1 scores, naïve Bayes classifiers exhibit the best performance, closely followed by logistic regression and SVM. Decision trees reveal slightly lower performance, falling below 0.5 F-score values. However, decision trees outperform all models in terms of precision, which in practical terms means that a coding support system based on decision trees would be less likely to inadequately suggest a given code for an episode, i.e., a codes suggested by this method would be more likely to be correct. On the other hand, recall values are higher for logistic regression and SVM, with decision trees and Bayesian classifiers falling behind in this metric. Low recall values would, in practice, be traduced in a support system being more likely to miss ICD codes that should have been be assigned to a given episode (i.e. a larger proportion of correct codes would be overlooked). The higher recall rates observed for naïve Bayes and logistic regression models are naturally due to the selection of very low classification thresholds as yielding the best F1 score for each code.

We also analyzed performance values obtained for each code, whose results are found in Figure 6. An interesting observation is that model performance (for instance, in terms of F1 score) does not decrease steadily with code frequency. For example, code 595.0 (acute cystitis) shows better model performance for decision trees than the most frequent code (401.9 – hypertension, not otherwise specified), despite having much lower relative frequency (i.e., lower representativeness of the positive class and thereby much higher class imbalance). On the other hand, performance variations across codes do not exhibit the same pattern for the 4 tested supervised learning methods. Using FCBF, model performance as seen in Figure 6 still varies considerably across codes. Decision trees are mostly dominant in terms of precision, while recall is now considerably higher for SVM. Furthermore, it is also possible to observe that F1 scores obtained with the 4 methods tend to be similar for each ICD code, which is an interesting result since each method adopts a specific approach for modeling trends in data and making predictions on future unseen instances. In effect, the fact that performance is similar for different methods may indicate that the causes explaining variations in performance across codes reside not only in the prediction method itself, but especially in the clinical concepts and contents underlying each ICD code and in the type of information (features) used as groundwork

for model training, as well as the quality of the data itself collected in real-world settings. Following this rationale, we looked into some illustrative examples of ICD codes and corresponding features selected by the FCBF method.

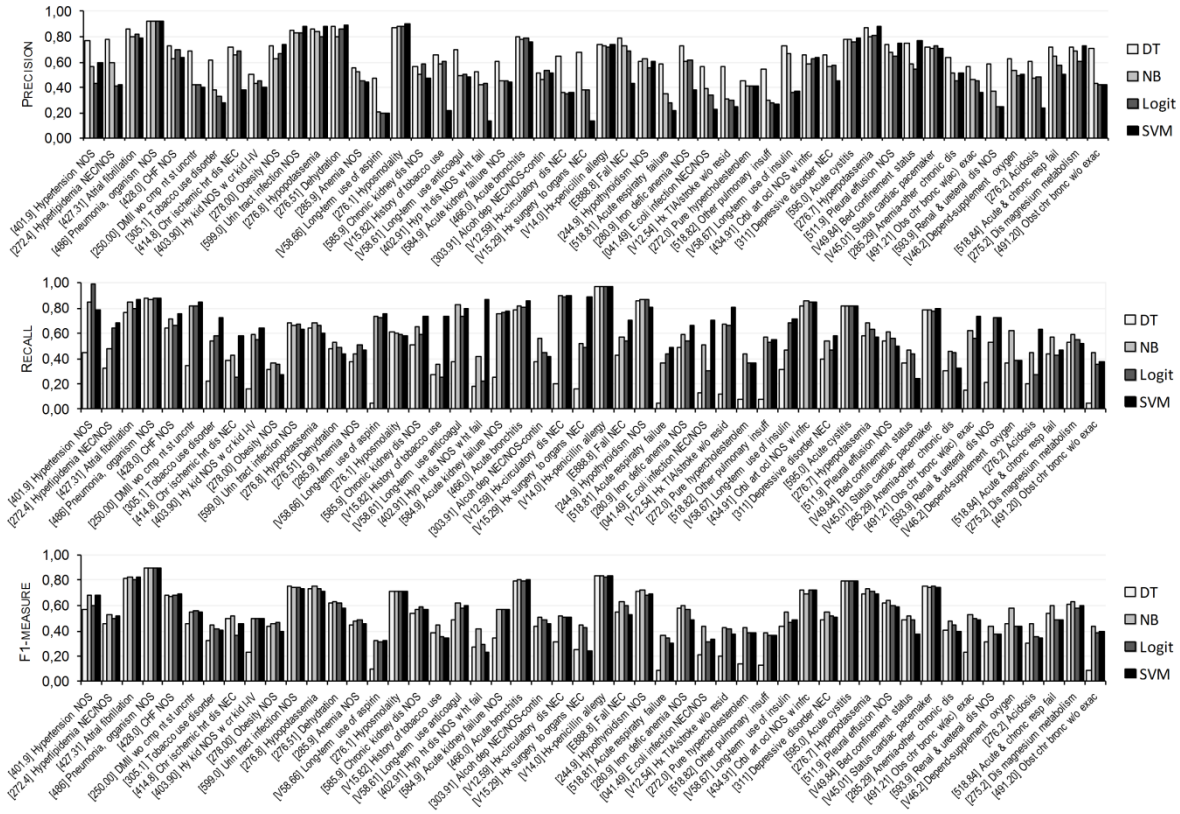

Figure 6 – Precision, recall and F1 scores obtained with decision trees, naïve Bayes, logistic regression and support vector machine models using FCBF feature selection for the 50 most frequent ICD codes.

Table 2 – Examples of features selected with the FCBF method (dx – assigned diagnosis; med – prescribed medication; lab – laboratory test; DM – diabetes mellitus; NOS – not otherwise specified; NEC – not elsewhere classified).

Code	**401.9** Hypertension NOS	**280.9** Iron deficiency anemia NOS	**250.00** DMII w/o complication not stated uncontrolled	**276.8** Hypopotassemia
FCBF Selected features (examples)	Essential hypertension (dx)	Iron deficiency anemias (dx)	Rapid-acting insulin (med)	Hypopotassemia (dx)
	Chronic kidney disease (dx)	Trivalent iron (med)	DM II or unspecified type without complication (dx)	Potassium chloride (med)
	Age	Unspecified iron deficiency anemia (dx)	D.M. without complication (dx)	Electrolyte and fluid disorders NEC (med)
	Hypertensive heart disease with heart failure (dx)	Coombs test (lab)	Ocreotide (med)	Plasmatic osmolality (lab)
	Losartan (med)	Other and unspecified anemias (dx)	Unspecified disorder of metabolism (dx)	Retention of urine (dx)
	Aldosterone (lab)	Unspecified disorders of arteries and arterioles (dx)	Secondary DM without complication (dx)	Chronic kidney disease (dx)

Exemplificative features selected by the FCBF method in 4 illustrative ICD codes (shown in

Table 2) refer mostly to diagnoses, medication and prescriptions. The examples shown here are correlated at different levels with the corresponding ICD code whose assignment they were used to predict, referring either to the same or related clinical conditions, to medication prescribed for such conditions, or tests often performed in such contexts. Other features identified by the FCBF method were, in turn, not apparently related to the ICD code for which they were selected, which might have been due to additional statistical effects captured by the measurements inherent to FCBF calculations. Still, observing that feature sets selected by FCBF can be judged, according to domain knowledge, at least partially relevant and pertinent. This indicates that the chosen feature selection method may be suitable in this context.

Another important observation concerns the appearance of similar designations of diagnoses with different specificity levels. This situation arises in contexts wherein the level of diagnosis detail is not uniformly used by health professionals. Therefore, it may occur that two patients with the exact same condition may have their clinical record filled in differently (e.g. two patients with pure hypercholesterolemia may be diagnosed as having that specific condition or just as having a disorder of the lipoid metabolism). Despite not being clinically incorrect, this situation might have negative impact on our methodology for introducing dispersion of similar instances among feature values. Therefore, to investigate whether the granularity used upon assigning diagnoses had significant effect on model performance, we modified the original feature set by collapsing all features referring to assigned diagnoses (since all diagnoses assigned during care provision are mapped to ICD-9-CM beforehand, these features were collapsed to the category level). The approach in this experiment was identical in terms of parameter adjustment and stepwise use of selected features.

Table 3 – Performance variations obtained after collapsing features referring to diagnoses assigned during patient episodes.

	401.9			**280.9**			**250.00**			**276.8**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DT	-0,003	0,119	**0,086**	-0,046	0,035	**0,008**	0,009	-0,241	**-0,281**	-0,171	-0,001	**-0,071**
NB	0,000	0,031	**0,010**	-0,004	-0,005	**-0,005**	-0,007	-0,064	**-0,021**	-0,159	-0,028	**-0,085**
Logit	0,149	-0,163	**0,079**	0,041	-0,015	**0,007**	-0,013	-0,061	**-0,026**	-0,121	-0,044	**-0,077**
SVM	0,047	-0,065	**0,002**	0,291	-0,143	**0,101**	-0,012	-0,039	**-0,019**	-0,550	0,328	**-0,227**

Upon modifying the features related to diagnoses assigned by health professionals during care provision, model performance revealed changes as illustrated in

Table 3 for the same ICD codes. While in the two first selected ICD codes (hypertension and iron-deficiency anemia), F1 scores slightly increased for most methods, these figures decreased in the other two ICD codes (diabetes mellitus and hypopotassemia). These results indicate that the degree of granularity used upon defining features – in this particular case, diagnoses assigned during episodes – plays an important role in the predictive power of developed models, which may benefit from being adjusted for different ICD codes. More so, this adjustment is relevant considering the variable heterogeneity of clinical conditions amongst ICD code categories. Moreover, the impact of feature granularity is also related with factors associated with the variability across EHR system users, namely the variability in clinical data recording practices, i.e., the level of detail and comprehensiveness adopted upon producing clinical documentation. Such user-related factors are context-specific and depend on organizational specificities, not only related to configuration of each EHR system (namely, data documenting rules embedded within a system, such as requiring a certain level of detail upon assigning diagnoses during care provision), but also to training and data recording policies enforced amongst health professionals.

4.3.2. Influence of class imbalance

We also found relevant to investigate the applicability of the proposed methodology across the multitude of ICD codes occurring in real-world settings, considering the high imbalance of most codes, which is evidenced in

Table 4. This table shows the number J of ICD codes (ordered by decreasing number of occurrences) representing different percentages of the 39273 total code occurrences in our dataset, as well as the number of positive examples of the J^{th} code. To this end, we performed an experiment consisting in training models for a wider range of ICD codes (while the above experiments were focused on the 50 most frequent codes) and observing the variation in overall performance results. In this experiment, we tested the same 4 methods and used the same parameter adjustment approaches for each of them. Models were trained using all features

selected by the FCBF method for each ICD code, without the stepwise feature set growth used in prior experiments.

Table 4 – Relationship between the number of cumulative code occurrences, the corresponding number of most frequent codes (*J*) and the number of positive examples of the J^{th} code.

Cumulative occurrences	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
# Top codes (*J*)	1	4	9	19	35	59	98	161	272	544	2272
# positive ex.	2043	885	492	324	200	133	79	49	25	8	1

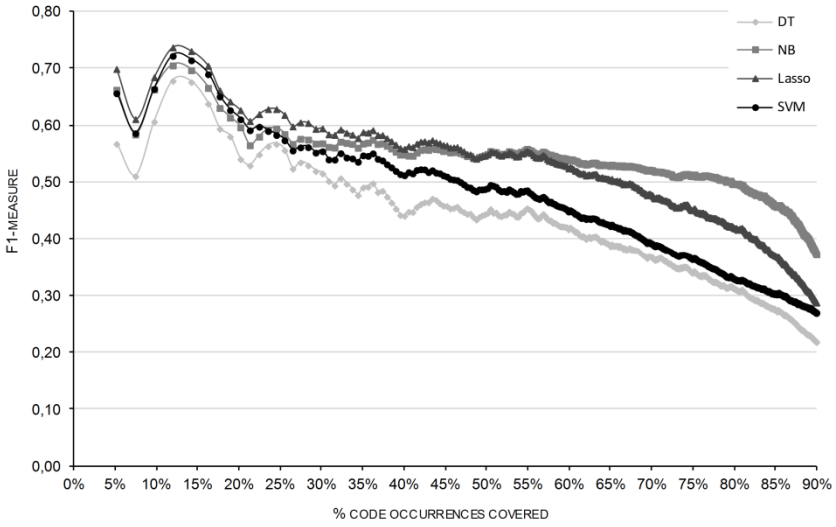

Figure 7 – Macro-averaged F1 scores as the percentage of covered ICD code occurrences is increased.

In this experiment, we considered F1 scores averaged between 5% and 90% of code occurrences (i.e. between 1 and 544 codes). The average performance decreases as more ICD codes are covered by the model training process and evaluated through cross-validation, which indicates that performance is increasingly lower for highly imbalanced codes, especially when covering more than 60% of total occurrences. Although in previous sections there was not an evident tendency of decreasing performance as the proportion of positive examples decreased (since the 50^{th} code still occurred 160 times), this tendency is revealed for more imbalanced codes, which indicates that an approach based on supervised learning models requires a certain representativeness of both positive and negative classes in order to be able to capture tendencies in data. More so, in datasets with high dimensionality, the use of large feature sets may also give rise to statistical artifacts

that additionally hamper prediction results, especially when using measures related with entropy, where one or some variables may discriminate the training set artificially well but not be adequate in the test set, thereby producing poor predictive performance. This also implies that feature selection itself might also suffer from such data artifacts. This analysis may be regarded as an indicator of the scope of applicability of the proposed methodology, especially showing for which ICD codes it might work better and signaling which relative frequencies are critically low, requiring further techniques to mitigate the negative influence of class imbalance.

5. Discussion and conclusions

In this paper, we propose a methodology to develop models to predict the assignment of ICD-9-CM codes using fully structured (and routinely collected) EHR data. The methodology makes use of a feature set based on the data model of a real-world EHR system, applies a filter method (FCBF) to perform feature selection and mitigate issues arising with high data dimensionality, and tests four different supervised learning models to assess their capability to predict code assignment. To tackle the multi-label nature of clinical coding, a binary relevance technique was employed to transform the original multi-label problem into multiple binary classification problems. The results of a set of experiments performed with a dataset of episodes from inpatient medical wards showed, firstly, the importance of using feature selection to reduce the dimensionality of the datasets prior to training models. Performance values (in terms of precision, recall and F1 scores) obtained with the different methods varied considerably across codes, and codes with reasonable prevalence did not exhibit an evident tendency of decreasing performance with decreasing prevalence. Moreover, performance variations across codes were approximately similar with the four tested methods, suggesting that performance is more likely to be related with the type of ICD code being predicted than with the specific model employed for this purpose. An overall analysis of results shows that while it is possible to predict well the assignment of certain codes, there are other codes for which models are not able to achieve good predictive performance. These results provide insight on the possibility of using fully structured data to support clinical coding, suggesting that such approach may have potential to effectively coding assist, and that

further methodological adaptations are required to improve the overall precision and recall figures yielded by the different methods.

The use of structured data is considered relevant in light of the evolution of EHR systems towards more structured data formats, by modifying data capturing mechanisms and enabling more complex data analyses with interest to the management of health organizations. The use of structured data circumvents the need to use NLP to extract concepts from free text, which proves challenging in many contexts where fewer NLP tools are available (e.g. for a specific language and domain) and the quality and ambiguity of clinical texts hamper information extraction. Naturally, the adoption of a methodology based on structured data depends on the architecture and configuration of each EHR system, which varies across organizations. Nevertheless, the approach hereby proposed makes use of data elements which are typically transversal to most EHR systems, particularly in what concerns diagnoses, medications and other prescriptions, as well as allergies and personal history. The underpinning lies in properly processing data to minimize redundancies and unnecessary detail, and carefully defining features in terms of their type (distinguishing numerical from nominal) and their missing patterns. It is also crucial to ensure that the methodological approach used to model EHR data is adaptable to accommodate changes in a streamlined fashion, which is implemented in our methodology through an automated data processing framework using simplification maps and a master file where features are defined in terms of their type and the data fields to be inspected in order to assess their value for each instance. With this approach, the effort lies in initially creating the data processing mechanism, which requires identifying and structuring all data elements, mapping data fields to features and defining variable types and missing assumptions. However, necessary further adaptations in line with modifications in the EHR system are achieved with marginal tuning in the data processing framework.

Feature selection was found to be extremely relevant to tackle the issues of high dimensionality, more so when in several cases there is a limited number of training instances in relation to the number of features. In effect, high dimensionality is bound to arise when defining features based on structured data, especially when considering data elements selected from catalogs (in which a binary feature is considered for each item) and features with multiple categorical values, from which several dummy variables are derived. In this work, feature selection greatly improved the results obtained with the four supervise learning models, while also

contributing significantly to reduce the computation effort of tuning model-specific parameters and performing multiple experiments and to improve model interpretability, particularly in the case of decision trees, naïve Bayes and logistic regression. Given the large volumes of episodes and information comprised in such EHR systems, feature selection not only helped mitigating model overfitting and artifacts arising in this context, but, more importantly, played a central role in narrowing down such large volumes to intelligible and computationally feasible levels. Naturally, a critical sense is also crucial in this methodology, which may benefit from the introduction of domain knowledge to revise and complement the results of feature selection methods.

In this paper, we tested multiple supervised models based on rather different approaches to model trends in training data and extrapolate them to new unseen episodes, since this approach suits the rationale of using knowledge from past episodes to support coding in future episodes. Naïve Bayes and logistic regression models exhibited the best overall performance in terms of F1 scores, with the former exhibiting the best figures. However, naïve Bayes entails a potential drawback related to the need to estimate class-conditional probabilities, which in contexts with high dimensionality and limited training data may prevent the coverage of all possible combinations of feature values with outcome class. Therefore, we advocate that it may be preferable to employ logistic regression models because of their ability to circumvent this issue, while still allowing reasonable model interpretability in terms of estimated coefficients and outcome probabilities (in this case, using a discriminative approach). SVM and decision trees exhibited lower performance values, yet both with their strengths and weaknesses. SVM, on the one hand, have the highest recall rate (i.e., are less prone to overlooking codes), but exhibit lower precision and consist of a black box with poor interpretability of results. Decision trees, on the other hand, have the best precision values (i.e., are less prone to suggest inadequate codes), but have poor recall values and are much more likely to miss codes, which is also not desirable in a coding support system. Naturally, the results obtained with the four methods also depend on intrinsic characteristics of the hospital organizational environment and processes, and consequently on the dataset itself. There is still a debate on how to evaluate and choose coding support methods and whether to use straightforward or more complex measures considering usability [67]. In effect, the choice of a method always involves a trade-off between aspects related to performance measures, their interpretability and

scalability, which are weighed according to the particular context and the interests of the different stakeholders involved in the development and use of a coding support system, including coding and clinical professionals, healthcare managers and administrative staff.

The methodology proposed in this paper consists of several building blocks for which future potential developments were identified. Firstly, the definition of a feature set based on the EHR structure may benefit from introducing domain knowledge from coding experts, namely in revising the way clinical and administrative data are modeled in order to further mitigate redundancy and unnecessary detail, since in this work we used a simpler, straightforward approach of listing data fields and defining features. On the other hand, not only the modeling of EHR data structure is important, but data quality is also crucial to properly use historical data. Poor data quality may be reflected in erroneous, absent or heterogeneously granular data, which have potentially negative impact on results. It is therefore important to assess correctness and completeness of clinical records used for model training and identify critical issues with impact in model performance. Furthermore, it is also important to implement strategies to improve data quality and completeness by providing training to health professionals on data recording policies, introducing modifications in EHR system configuration to prevent introduction of erroneous and ambiguous data, and lastly capitalize additional data elements such as laboratory and other exam results to fill in feature values to increase the odds of all relevant findings being captured by feature values. Such developments on data structuring and modeling stages entail potential impacts not only on the EHR system and data quality itself, but at the organizational level modifying practices and guidelines for health professionals.

In data processing and model development stages, several lines of improvement are also relevant. The approach to handle missing data – either deleting, imputing or classifying – is a recurrent issue and is likely to have implication in model results [68]. In this paper, features with missing values were left out of model development, yet some of these features, such as blood pressure values, may have contained data with relevance to predict the assignment of certain ICD codes (e.g. hypertension-related codes), and it appears relevant to investigate the use of missing data. Additionally, data imbalance has showed to hamper model performance for rare codes, as would be expected using a supervised learning approach, and several approaches have been proposed in the literature [59]. As for feature selection, in turn, since it is also a field of

study in itself, may benefit from testing alternative approaches which are able to handle large datasets amongst the myriad of new methods constantly being developed [69]. Moreover, we also intend to introduce domain knowledge, with input from experts in revising feature selection to complement algorithmic feature selection methods, not only aiming for performance improvement, but also to contributing to model interpretability.

Clinical coding has long posed a central issue in health organizations. Its automation has been approached numerous times through different techniques, most often based on free text processing to extract knowledge of which to take advantage in predicting code assignment. To the best of our knowledge, our work is pioneer in proposing a methodology to provide coding support based on fully structured data. While the issues of processing free text are avoided through this approach, other challenges naturally arise, especially when dealing with large data volumes, complex EHR systems and different degrees of system implementation maturity and user proficiency. The results shown in this paper represent a comprehensive first approach (setting off from structured EHR data in its raw format and establishing a pipeline to build supervised models for making predictions) shedding light onto the feasibility of such approach. Although still demanding further improvements in predictive performance to allow implementation in real settings, we identified several lines of improvement on which we intend to capitalize, also considering the eventual evolution of EHR system technology and associated organizational changes.

6. Acknowledgments

The authors acknowledge the support from Fundação para a Ciência e a Tecnologia (SFRH/BDE/51605/2011), from Siemens SA and from the Centre for Management Studies of Instituto Superior Técnico (CEG-IST). The authors would also like to thank Hospital Professor Doutor Fernando Fonseca for close collaboration and availability throughout this research.

7. Conflicts of interest

The authors state no conflicts of interest.

8. References

[1] L. A. Schraffenberger, *Basic ICD-9-CM Coding 2006 Edition*. AHIMA, 2006.

[2] M. J. Bowie and R. M. Schaffer, *Understanding ICD-9-CM Coding: A Worktext*. 2011.

[3] Accountability Act, "ICD-9-CM Official Guidelines for Coding and Reporting," pp. 1–107, 2011.

[4] AHIMA computer-assisted coding e-HIM work group, "Delving into Computer-assisted Coding (AHIMA Practice Brief)," *J. AHIMA*, vol. 75, p. 48A – 48H, 2004.

[5] V. Patel, E. Jamoom, C.-J. Hsiao, M. Furukawa, and M. Buntin, "Variation in Electronic Health Record Adoption and Readiness for Meaningful Use: 2008–2011," *J. Gen. Intern. Med.*, pp. 1–8, 2013.

[6] E. W. Ford, N. Menachemi, and T. Phillips, "Predicting the Adoption of Electronic Health Records by Physicians : When Will Health Care be Paperless ?," no. 13, pp. 106–113, 2006.

[7] J. G. Anderson, "Social, ethical and legal barriers to e-health.," *Int. J. Med. Inform.*, vol. 76, no. 5–6, pp. 480–3, 2006.

[8] J. D. Hatton, T. M. Schmidt, and J. Jelen, "Adoption of Electronic Health Care Records: Physician Heuristics and Hesitancy," *Procedia Technol.*, vol. 5, pp. 706–715, Jan. 2012.

[9] R. H. Miller and I. Sim, "Physicians' Use Of Electronic Medical Records: Barriers And Solutions," *Health Aff.*, vol. 23, no. 2, pp. 116–126, Mar. 2004.

[10] P. E. J. E. Embi, T. H. R. Yckel, J. U. R. Logan, J. U. L. Bowen, T. H. G. Cooney, and P. N. Gorman, "Impacts of Computerized Physician Documentation in a Teaching Hospital: Perceptions of Faculty and Resident Physicians," *J. Am. Med. Informatics Assoc.*, vol. 11, pp. 300–309, 2004.

[11] C. M. Cusack, G. Hripcsak, M. Bloomrosen, S. T. Rosenbloom, C. a Weaver, A. Wright, D. K. Vawdrey, J. Walker, and L. Mamykina, "The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting.," *J. Am. Med. Informatics Assoc.*, vol. 20, pp. 134–40, Jan. 2013.

[12] J. J. Nadler and G. J. Downing, "Liberating Health Data for Clinical Research Applications," *Sci. Transl. Med.* , vol. 2 , no. 18 , pp. 18cm6–18cm6, Feb. 2010.

[13] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records.," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 1, pp. 117–21, Jan. 2013.

[14] G. Hripcsak, C. Knirsch, L. Zhou, A. Wilcox, and G. Melton, "Bias associated with mining electronic health records.," *J. Biomed. Discov. Collab.*, vol. 6, pp. 48–52, Jan. 2011.

[15] H. J. Tange, H. C. Schouten, A. D. M. Kester, and A. Hasman, "The granularity of medical narratives and its effect on the speed and completeness of information retrieval," *J. Am. Med. Inform. Assoc.*, vol. 5, no. 6, pp. 571–82, 1998.

[16] G. Hripcsak, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton, "Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing," *Ann. Intern. Med.*, vol. 122, no. 9, pp. 681–688, May 1995.

[17] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems.," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 6, pp. 646–51, 2010.

[18] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research.," *Yearb. Med. Inform.*, pp. 138 – 154, Jan. 2008.

[19] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner, "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions.," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 540–3, 2011.

[20] Q. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Med. Inform. Decis. Mak.*, vol. 6, no. 1, p. 30, 2006.

[21] M. L. Gundersen, P. J. Haug, T. a Pryor, R. van Bree, S. Koehler, K. Bauer, and B. Clemons, "Development and evaluation of a computerized admission diagnoses encoding system," *Comput. Biomed. Res.*, vol. 29, pp. 351–72, Oct. 1996.

[22] A. R. Aronson, O. Bodenreider, D. Demner-fushman, K. W. Fung, V. K. Lee, J. G. Mork, A. Névéol, L. Peters, and W. J. Rogers, "From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches," in *BioNLP 2007: Biological, translational, and clinical language processing*, 2007, no. June, pp. 105–112.

[23] I. Goldstein, A. Arzrumtsyan, and O. Uzuner, "Three approaches to automatic assignment of ICD-9-CM codes to radiology reports.," *AMIA Annu. Symp. Proc.*, pp. 279–83, Jan. 2007.

[24] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salantera, and T. Salakoski, "Machine Learning to Automate the Assignment of Diagnosis Codes to Free-text Radiology Reports: a Method Description," in *ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland*, 2008.

[25] Y. Zhang, "A Hierarchical Approach to Encoding Medical Concepts for Clinical Notes," in *Proceedings of the ACL-08: HLT Student Research Workshop*, 2008, no. June, pp. 67–72.

[26] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, "Automatic Code Assignment to Medical Text," in *BioNLP 2007: Biological, translational, and clinical language processing*, 2007, no. June, pp. 129–136.

[27] P. Matykiewicz, W. Duch, and J. Pestian, "Associating Medical Concept Relations with ICD-9-CM Coding Rules," 2006.

[28] R. Farkas and G. Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems," *BMC Bioinformatics*, vol. 9, no. Suppl 3, p. S10, Jan. 2008.

[29] D. Delamarre, a Burgun, L. P. Seka, and P. Le Beux, “Automated coding of patient discharge summaries using conceptual graphs.,” *Methods Inf. Med.*, vol. 34, no. 4, pp. 345–51, Sep. 1995.

[30] P. Franz, A. Zaiss, S. Schulz, U. Hahn, and R. Klar, “Automated coding of diagnoses - three methods compared.,” in *Proceedings of AMIA Annual Symposium*, 2000, pp. 250–4.

[31] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak, “Automated Encoding of Clinical Documents Based on Natural Language Processing,” *J. Am. Med. Informatics Assoc.*, pp. 392–402, 2004.

[32] L. Kevers and J. Medori, “Symbolic Classification Methods for Patient Discharge Summaries Encoding into ICD.”

[33] R. Kukafka, M. E. Bales, A. Burkhardt, and C. Friedman, “Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health,” *J. Am. Med. Informatics Assoc.*, vol. 13, no. 5, pp. 508–515, 2006.

[34] L. S. Larkey and W. B. Croft, “Automatic Assignment of ICD9 Codes To Discharge Summaries,” 1995.

[35] S.-T. Li and C.-C. Chen, “Conceptual-driven classification for coding advise in health insurance reimbursement,” *Artif. Intell. Med.*, vol. 51, no. 1, pp. 27–41, 2011.

[36] Y. A. Lussier, L. Shagina, and C. Friedman, “Automating ICD-9-CM Encoding Using Medical Language Processing: A Feasibility Study,” in *Proceedings of AMIA Annual Symposium*, 2000, vol. 1, p. 5027.

[37] Y. a Lussier, L. Shagina, and C. Friedman, “Automating SNOMED coding using medical language understanding: a feasibility study.,” *Proc. AMIA Annu. Symp.*, pp. 418–22, Jan. 2001.

[38] J. Medori and P. B. Pascal, “Machine learning and features selection for semi-automatic ICD-9-CM encoding,” in *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, 2010, no. June, pp. 84–89.

[39] H. P. Dinwoodie and R. W. Howell, “Automatic disease coding: the ‘fruit-machine’ method in general practice.,” *Br. J. Prev. Soc. Med.*, vol. 27, no. 1, pp. 59–62, Feb. 1973.

[40] L. V. Lita, S. Yu, S. Niculescu, and J. Bi, “Large Scale Diagnostic Code Classification for Medical Patient Records,” in *Proceedings of the International Joint Conference on Natural Language Processing*, 2008.

[41] W. C. Morris, D. T. Heinze, H. R. Warner Jr, A. Primack, a E. Morsch, R. E. Sheffer, M. a Jennings, M. L. Morsch, and M. a Jimmink, “Assessing the accuracy of an automated coding system in emergency medicine.,” in *Proceedings of AMIA Annual Symposium*, 2000, pp. 595–9.

[42] S. V. S. Pakhomov, J. D. Buntrock, and C. G. Chute, “Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques,” *J. Am. Med. Informatics Assoc.*, pp. 516–525, 2006.

[43] P. Ruch, J. Gobeilla, I. Tbahritia, and A. Geissbühlera, “From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding,” in *Proceedings of AMIA Annual Symposium*, 2008, pp. 636–40.

[44] S. Pereira, A. Névéol, P. Massari, M. Joubert, and S. Darmoni, “Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding.,” *Stud. Health Technol. Inform.*, vol. 124, pp. 845–50, Jan. 2006.

[45] Y. Yan, G. Fung, R. Rosales, and J. G. Dy, “Medical Coding Classification by Leveraging Inter-Code relationships,” in *Proceedings of the Annual SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010, pp. 193–202.

[46] J.-W. Xu, S. Yu, J. Bi, L. V. Lita, R. S. Niculescu, and R. B. Rao, “Automatic medical coding of patient records via weighted ridge regression,” in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, 2007, vol. 0, pp. 260–265.

[47] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, “A General Natural-Language Text Processor for Clinical Radiology,” *J. Am. Med. Informatics Assoc.*, vol. 1, pp. 161–174, 1994.

[48] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances.,” *J. Am. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–36, 2010.

[49] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, “A simple algorithm for identifying negated findings and diseases in discharge summaries.,” *J. Biomed. Inform.*, vol. 34, no. 5, pp. 301–10, Oct. 2001.

[50] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, “The Unified Medical Language System,” *Methods Inf. Med.*, vol. 32, no. 4, pp. 281–291, Aug. 1993.

[51] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, and N. Elhadad, “Diagnosis code assignment: models and evaluation metrics.,” *J. Am. Med. Inform. Assoc.*, pp. amiajnl–2013–002159, Dec. 2013.

[52] C. M. Bishop, *Pattern Recognition and Machine Learning*. Singapore: Springer, 2006.

[53] J. An, X. Lu, and H. Duan, “Integrated Visualization of Multi-Modal Electronic Health Record Data,” in *2nd International Conference on Bioinformatics and Biomedical Engineering*, 2008, pp. 640–643.

[54] R. Haux, C. Seggewies, W. Baldauf-Sobez, P. Kullmann, H. Reichert, L. Luedecke, and H. Seibold, “Soarian--workflow management applied for health care.,” *Methods Inf. Med.*, vol. 42, no. 1, pp. 25–36, Jan. 2003.

[55] G. Tsoumakas, I. Katakis, and I. Vlahavas, “A Review of Multi-Label Classification Methods.”

[56] T. Mitchell, *Machine Learning*, vol. 4, no. 1. McGraw-Hill, 1997, p. 432.

[57] L. Rokach and O. Maimon, “Classification Trees,” in in *Data Mining and Knowledge Discovery Handbook*, Second Edi., O. Maimon and L. Rokach, Eds. New York: Springer, 2010, pp. 149–174.

[58] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd Editio. John Wiley & Sons, Inc., 2000.

[59] H. He and E. A. Garcia, “Learning from Imbalanced Data,” vol. 21, no. 9, pp. 1263–1284, 2009.

[60] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[61] L. Yu and H. Liu, “Efficient Feature Selection via Analysis of Relevance and Redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[62] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[63] S. Wood, *Generalized Additive Models: An Introduction with R (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2006, p. 410.

[64] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining Multi-label Data,” in in *Data Mining and Knowledge Discovery Handbook*, O. Mainon and L. Rokach, Eds. New York, NY: Springer, 2010, pp. 667–685.

[65] H. Liu, F. Hussain, C. L. I. M. Tan, and M. Dash, “Discretization: An Enabling Technique,” *Data Min. Knowl. Discov.*, vol. 6, pp. 393–423, 2002.

[66] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, “Conditional Likelihood Maximisation : A Unifying Framework for Information Theoretic Feature Selection,” *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, 2012.

[67] J. Puentes, J. Montagner, L. Lecornu, and J.-M. Cauvin, “Information quality measurement of medical encoding support based on usability.,” *Comput. Methods Programs Biomed.*, vol. 112, no. 3, pp. 329–42, Dec. 2013.

[68] F. Cismondi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. C. Sousa, and S. N. Finkelstein, “Missing data in medical databases: impute, delete or classify?,” *Artif. Intell. Med.*, vol. 58, no. 1, pp. 63–72, May 2013.

[69] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, “Feature Selection : An Ever Evolving Frontier in Data Mining,” in *JMLR: Workshop and Conference Proceedings*, 2010, pp. 4–13.