Noname manuscript No.
(will be inserted by the editor)

Diagnosis Prediction from Electronic Health Records (EHR) using the Binary Diagnosis History Vector Representation

Ieva Vasiljeva · Ognjen Arandjelović*

School of Computer Science

University of St Andrews

St Andrews KY16 9SX

Fife, Scotland

United Kingdom

E-mail:

ognjen.arandjelovic@gmail.com

Tel: +44 (0)1334 46 28 24

Address(es) of author(s) should be given

Abstract Large amounts of rich, heterogeneous information nowadays routinely collected by health care providers across the world possess remarkable potential for the extraction of novel medical data and the assessment of different practices in real-world conditions. Specifically in this work our goal is to use Electronic Health Records (EHRs) to predict progression patterns of future diagnoses of ailments for a particular patient, given the patient's present diagnostic history. Following the highly promising results of a recently proposed approach which introduced the diagnosis history vector representation of a patient's diagnostic record, we introduce a series of improvements to the model and conduct thorough experiments that demonstrate its scalability, accuracy, and practicability in the clinical context. We show that the model is able to capture well the interaction between a large number of ailments which correspond to the most frequent diagnoses, show how the original learning framework can be adapted to increase its prediction specificity, and describe a principled, probabilistic method for incorporating explicit, human clinical knowledge to overcome semantic limitations of the raw EHR data.

Keywords Electronic medical records · EMRs · Bayesian · risk · disease · epidemiology

1 Introduction

The trend of increased efforts in health data collection and its ready digitization is widely recognized as a major change in the manner medical data is used. In particular the collection of Electronic Health Records (EHRs) has recently started attracting major translational research efforts in the domains of data mining, knowledge extraction, and machine learning (Xu et al., 2016; Christensen and Ellingsen, 2016). Electronic Health Records have already been extensively used in large scale sociodemographic surveys of death causes (RGI-CGHR Collaborators, 2009), clinical epidemiological (Paul et al., 2015c; Bhatnagar et al., 2015; Crawford et al., 2010) and pharmacoepidemiological studies (Wettermark et al., 2013; Lau et al., 2011; Paul et al., 2015b), as well as in the analysis of pharmacovigilance (Nadkarni, 2010; Liu et al., 2013; Coloma et al., 2013), health related economic effects (Canavan et al., 2015; Bessou et al., 2015), and public health (Birkhead et al., 2015; Paul et al., 2015a; Kukafka et al., 2007; Menachemi and Collum, 2011). Considering that this research is still in its early stages it is undeniably wise to refrain from overly ambitious predictions regarding the type of knowledge which may be discovered in this manner, at the very least it is true that few domains of application of the aforesaid techniques hold as much promise for impact. It is sufficient to observe the potential benefits that an increased understanding of complex interactions of lifestyle diseases in the economically developed world could deliver in terms of personalized medicine or health care policy (Fan et al., 2016) on the one hand, and a wiser utilization of resources, aid, and educational material in the economically deprived countries

(RGI-CGHR Collaborators, 2009), to appreciate the global and overarching potential.

Public health care is an issue of major global significance and concern. On one end of the spectrum, the developing world is still plagued by "diseases of poverty" which are nearly non-existent in the most technologically developed countries; on the other end, the health risk profile of industrially leading nations has dramatically changed in recent history with an increased skew towards so-called "diseases of affluence", as illustrated in Figure 1 (data taken from (Murray et al., 2001)).

Hence, health care management poses challenges both in the sphere of policy making and scientific research. Considering the complexity of problems at hand, it is unsurprising that there is an ever-increasing effort invested in a diverse range of promising avenues. Yet, the available resources are inherently limited. To ensure their best usage it is crucial both to develop an understanding of the related epidemiology, as well as to be able to communicate this knowledge effectively to those who can benefit from it: governments (Berwick and Hackbarth, 2012), the medical research community (Beykikhoshk et al., 2015a, 2016; Andrei and Arandjelović, 2016), health care practitioners (Arandjelović, 2015a; Osuala and Arandjelović, 2017), and patients (Beykikhoshk et al., 2014; Barracliffe et al., 2017).

The associations between diseases and a wide variety of risk factors are underlain by a complex web of interactions. This is particularly the case for the diseases of the developed world. The key premise of the present work is that to facilitate the understanding of this complexity and the discovery of meaningful

patterns within it, it is crucial to make use of the vast amounts of data routinely collected by health services in industrially and technologically developed countries.

Our specific aim is to develop a framework which allows a health practitioner (e.g. a doctor or a clinician) to manipulate the available patient information in an intuitive yet powerful fashion. Such a framework would, on one end of the utility spectrum, facilitate a deepening of disease understanding, and on the other, provide the practitioner with a tool which can be used to incentivize the patient at risk to make the required lifestyle changes.

1.1 Data: electronic medical records

This work leverages the large amounts of medical data routinely collected and stored in electronic form by health providers in most developed countries. This is a rich data source which contains a variety of information about each patient including the patient's age and sex, mother tongue, religion, marital status, profession, etc. In the context of the present work, of main interest is the information collected each time a patient is admitted to the hospital (including out-patient visits to general practitioners or specialists). The format of this data is explained next.

Each time a patient is admitted to the hospital the reason for the admission, as determined by the medical practitioner in primary charge during the admission, is recorded in the patient's medical history. This is performed using a standardized coding schema such as that provided by the International Statistical Classi-

fication of Diseases and Related Health Problems (ICD-10) (World Health Organization, 2004) and the related Australian Refined Diagnosis-Related Groups (AR-DRGs).

These have hierarchical structures (Arandjelović, 2016). ICD-10, for example, contains 22 chapters, each chapter encompassing a spectrum of related health issues (usually symptomatically rather than etiologically related). For example, ICD-10 Chapter 4 which includes codes E00-E90, covers "Endocrine, nutritional and metabolic diseases". At each subsequent depth level of the tree the grouping is refined and the scope of conditions narrowed down. In this paper we use the classification attained at the depth of two of ICD-10, which achieves a good compromise between specificity and frequency of occurrence. This results in each diagnosis being given a three character code which comprises a leading capital letter (A-Z, first grouping level), followed by a two digit number (further refinement). For example, E66 codes for "Obesity" within the broader range of "Endocrine, nutritional and metabolic diseases".

2 Modelling comorbidity progression

The major contribution of this work is a novel disease progression model. The principal challenge is posed by the need for a model which is sufficiently flexible to be able to capture complex patterns of comorbidity development, while at the same time constrained enough to facilitate learning from a real-world data corpus.

2.1 Bottom-up modelling

The problem of modelling disease progression has already attracted a considerable amount of research attention. Most previous research focuses on specific individual diseases, such as type-II diabetes mellitus (Topp et al., 2000; De Gaetano et al., 2008) or heart disease (Ye et al., 2012). These methods are inherently 'low-level' based in the sense that they explicitly model known physiological changes that affect disease progression. For example, the modelling of the progression of type-II diabetes may include low-level models of β-cell mass changes, and insulin and glucose dynamics (Topp et al., 2000), with the free parameters (e.g. β-cell replication rate) of the models adopted from previous empirical studies. Higher level disease progression then emerges from the interaction of low-level models.

The low-level approach to disease modelling has several limitations. Firstly, by their very nature these models are limited to specific diseases only and cannot be readily adapted to deal with conditions with entirely different etiologies. Secondly, the modelling is practically constrained usually to a single condition, two at the most, as the complexity of modelled system increases dramatically with the inclusion of a greater number of conditions. This observation is of major significance as most diseases of the developed world are most often accompanied and affected by multiple comorbidities. Lastly, the range of diseases which can be modelled in this manner is limited to diseases which are sufficiently well understood and studied to allow for the free model parameters to be set reliably; even for type-II diabetes, which has been studied extensively, at present some

parameters must be set in an *ad hoc* manner and others using *in vitro* rather than *in vivo* data (Topp et al., 2000).

2.2 Direct high-level modelling

Given the significance of the disadvantages of low-level based disease progression models, in this paper an alternative approach is pursued, that of seeking to describe disease progression as well as the interplay of different comorbidities directly on the 'high-level' as observed by a medical practitioner. Previous research in this area is far scarcer than that on low-level modelling; a possible reason for this is probably to be found in the until recently limited availability of large-scale medical records data. The central idea of the existing corpus of work is to regard disease progression as a discrete sequence of events, with the progression governed by what is assumed to be a first-order Markov process (Sukkar et al., 2012; Jackson et al., 2003).

A high-level view of disease progression is seen as being reflected by a patient's diagnostic history $H = d_1 \rightarrow d_2 \rightarrow \ldots \rightarrow d_n$ where d_i is a discrete variable whose value is a code corresponding to the i-th of n diagnoses on the patient's record. The parameters of the underlying first-order Markov model are then learnt by estimating transition probabilities $p(d' \rightarrow d'')$ for all transitions encountered in training (the remaining transition probabilities are usually set to some low value rather than 0, using a pseudocount based estimate) (Wang et al., 2014; Folino and Pizzuti, 2011; Bartolomeo et al., 2008). The model can be applied to predict the diagnosis d_{n+1} expected to follow from the current history

by model likelihood maximization:

$$d_{n+1} = \arg\max_{d} p(d_n \to d). \tag{1}$$

Alternatively, it may be used to estimate the probability of a particular diagnosis d^* at some point in future:

$$p_f(d^*) = \sum_{d} [p(d \to d^*)\, p_f(d)], \tag{2}$$

or to sample the space of possible histories:

$$H' = d_1 \to d_2 \to \ldots \to d_n \dashrightarrow d_{n+1} \dashrightarrow d_{n+2} \ldots. \tag{3}$$

The primary purpose of the Markovian assumption is to constrain the mechanism underlying a specific process and thus formulate it in a manner which leads to a tractable learning problem. Although it is seldom strictly true, that it is often a reasonable approximation to make is witnessed by its successful application across a diverse range of disciplines; examples of modelled phenomena include meteorological events (Gabriel and Neumann, 1962), software usage patterns (Whittaker and Thomason, 1994), breast cancer screening (Duffy and Yau, 1995), human motion and behaviour (Lee et al., 2005; Arandjelović, 2011), and many others. Nonetheless, the key premise motivating the model in this paper is that the Markovian assumption is in fact not appropriate for the high-level modelling of disease progression (note that this does not reject its possible applicability in disease progression modelling on different levels of abstraction). Indeed, we will demonstrate this empirically. The aforementioned premise is readily substantiated using a theoretical argument as well. Consider a patient who is admitted for what is diagnosed as a serious chronic illness. If the

same patient is subsequently admitted for an unrelated ailment, possibly a trivial one, the knowledge of the serious underlying problem is lost and the power to predict the next related diagnosis lost. The model proposed in the section which follows solves this problem, while at the same retaining the tractability of Markov process based approaches.

2.3 Proposed approach

In this paper our aim is to predict the probability of a specific diagnosis a following the patient history H:

$$p(H \rightarrow a|H). \tag{4}$$

The difficulty of formulating this as a tractable learning problem lies in the fact that the space of possible histories is infinite as H can be of an arbitrary length. Even if the length $l(H)$ is limited, the number of possible histories is extremely large: $[l(H)]^{n_a}$ where n_a is the number of different diagnosis codes. Therefore it is necessary to make an approximation which constrains and simplifies the task. We already argued why the Markovian assumption on the level of diagnosis codes is inappropriate. In its stead we propose a different representation of a patient's state, particularly suitable for the modelling of disease progression (Arandjelović, 2015b). Consider a particular diagnosis history $H = d_1 \rightarrow \ldots \rightarrow d_n$. The proposed method makes use of the well known observation that when it comes to chronic diseases, the very *presence* of past complications strongly predicts future complications (Mudge et al., 2011; Friedman et al., 2008–2009; Dharmarajan et al., 2013; Butler and Kalogeropoulos, 2012).

Thus, a history H is represented using a history vector $v = v(H)$ which is a fixed length vector with binary values (Beykikhoshk et al., 2015b). Each vector element corresponds to a specific diagnosis code (except for one special element explained shortly) and its value is 1 if and only if the corresponding diagnosis is present in the history:

$$\forall d \in D.\ v(H)_{i(d)} = \begin{cases} 1 : \exists j.\ H = H_1 \rightarrow d_j \rightarrow H_2 \wedge d = d_j \\ 0 : \text{otherwise} \end{cases}$$

where D is the set of diagnosis codes, $i(d)$ indexes the diagnosis code d in a history vector, and $H_{1,2}$ may take on degenerate forms of empty histories. By collapsing an arbitrary length history of diagnoses onto a fixed length vector, the space of possible states over which learning is performed is dramatically reduced and the problem immediately made far more tractable. Notice the importance of the observation that it is the *presence* of past complications which most strongly predicts future ailments, given that under this representation any information on the ordering of diagnoses is discarded. The binary nature of the representation also has the effect of reducing the size of the space over which inference is performed. In this case, this is achieved by discarding information on the number of repeated diagnoses and in this manner it too predicates the overwhelming predictive power of the presence of history of a particular ailment, rather than the number of the corresponding diagnoses.

The disease progression modelling problem at hand is thus reduced to the task of learning transition probabilities between different patient history vectors:

$$p(v(H) \rightarrow v(H')). \tag{5}$$

It is important to observe that unlike in the case of Markov process models working on the diagnosis level when the number of possible transition probabilities is close to ${n_a}^2$, here the transition space is far sparser. Specifically, note that it is impossible to observe a transition from a history vector which codes for the existence of a particular past diagnosis to one which does not, that is:

$$v(H)_{i(d)} = 1 \wedge v(H')_{i(d)} = 0 \Rightarrow p(v(H) \rightarrow v(H')) = 0. \tag{6}$$

The converse does not hold however. Moreover, possible transitions can be only those which include either no changes to the history vector (repeated diagnosis) or which encode exactly one additional diagnosis:

$$p(v(H) \rightarrow v(H')) \begin{cases} > 0 : \forall a.\, v(H)_{i(d)} = 1 \Rightarrow v(H')_{i(d)} = 1 \\ \qquad \text{and} \\ \quad \big|\{a : v(H)_{i(d)} = 1\}\big| \leq 1 + \big|\{a : v(H')_{i(d)} = 1\}\big| \\ \\ = 0 : \text{otherwise} \end{cases} \tag{7}$$

This gives the upper bound for the number of non-zero probability transitions of $n_a \times 2^{n_a}$. In practice the actual number of transitions is far smaller (several orders of magnitude for the data set described in the next section) which allows the learnt model to be stored and accessed efficiently.

The final aspect of the proposed model concerns transitions with probabilities which do not vanish but which are nonetheless very low. These transitions can be reasonably considered to be noise in the sense that the corresponding

probability estimates are unreliable due to low sample size. Hence diagnosis history vectors are constructed using only the $\hat{n}_d$ most common diagnoses and merge the remaining $n_d - \hat{n}_d$ types into a single special code 'other'. Thus, the dimensionality of diagnosis history vectors becomes $\hat{n}_d + 1$. The soundness of this approach can be readily observed by examining the plot in Figure 2 which shows that only a small number of diagnosis types covers a vast number of all data. For example the top 30 most frequent types account for 75% of all diagnoses.

A conceptual illustration of the method is shown in Figure 3.

2.4 Limitations and questions

One of our contributions of the present work is in the form of an analysis which scrutinizes the expectation that the method would scale well. In the original work (Arandjelović, 2015b) it was argued that the predictive performance of the method, reported with explicit modelling of the 30 most frequent diagnosis types only, could be maintained as a greater number of diagnosis types is included in the model as most practical applications would demand. The original paper did not investigate this; rather, the number of salient, explicitly modelled diagnoses was set in an *ad hoc* manner to 30, explaining approximately 75% of the data corpus (Arandjelović, 2015b). If our expectation of performance deterioration with an increased number of explicitly modelled diagnoses is correct, and if the rate of deterioration is high, the model could end up being of little practical significance: on the one end of the parameter spectrum the model

would provide high accuracy but insufficient specificity for its predictions to be practically useful, and on the other high specificity but poor accuracy for its predictions to be relied upon. Thus an analysis of this aspect of the original method is necessary before any practical use can be considered; our experiments as regards this issue are presented in Section 4.3.

3 Further technical contributions

In this section we introduce our two main technical contributions. Our third contribution in the form of novel analyses and empirical results which highlight important and promising future research directions is presented in Section 4.

3.1 Improving the specificity of the model

The first major contribution of the present work goes to the very heart of the learning framework underlying the diagnostic progression model, and concerns the issue of the space over which learning is performed. In other words we propose a paradigm change in terms of what is explicitly learnt.

Recall from the previous section that the method described by (Arandjelović, 2015b) learns the probabilities of transitions from the space of history vectors to the same space of history vectors i.e. it learns $p(H'|H)$ where H is a patient history vector and H a possible extension to that history, $H' = H \rightarrow d$. This approach naturally follows from the structure of the problem: both H and H' are states in a Markov chain and indeed the baseline formulation of this class of problems learns amongst other things precisely these transition probabilities.

However, the very aspect of the history vector representation which makes it a powerful feature for longitudinal pattern extraction, in this instance introduces a significant practical limitation. Because history vectors are binarized, in general a specific transition does not uniquely determine the diagnosis which caused the transition to occur. In particular this occurs when a diagnosis already recorded in a patient's history is repeated – the transition from H to itself does not allow the method to distinguish between different diagnoses in the patient's history and determine which effected the transition (Vasiljeva and Arandjelović, 2016b). This is a major limitation given that many of the most serious diseases tend to be chronic in nature.

The method introduced in the present paper solves the described problem by changing the space over which learning is performed. In particular, rather than learning the probabilities of transitions between history vectors themselves, we learn the probabilities of follow up diagnoses directly. It can be readily seen that this is a stronger learning task in the sense that knowing the follow-up diagnosis d allows for the computation of the next Markov chain state $H' = H \rightarrow d$ without ambiguity whereas the opposite is not the case, as described previously. What makes this learning choice particularly sensible is that it does not carry the burden of either greater computational complexity nor learning challenge – the dimensionality of the space over which learning is performed stays exactly the same (it is governed by the choice of the number of salient diagnoses), which remains as densely populated as before. Hence this learning paradigm change is unambiguously superior to that described originally.

3.2 Risk driven inference

Our second key technical novelty concerns a major challenge in the development of models underlain by data from EHRs, which emerges from the pervasive problem known as the *semantic gap* (Vasiljeva and Arandjelović, 2016c). In colloquial terms, the problem is readily understood as arising from the lack of understanding of, say, disease aetiology and physiology that an automatic method has in the interpretation of data from EHRs. For example, a human expert (such as a general practitioner or a specialist) who does have such knowledge, may be readily able to discount even the consideration of certain disease interactions which may be difficult to infer using a purely data driven approach that machine methods generally employ. To overcome this challenge some means of interaction, that is, information provision between an expert and a computer algorithm is needed. Yet this interaction has to be intuitive, and require little effort and computing expertize.

The original authors correctly point out and thereafter empirically demonstrate that a major limitation in the use of Markovian models lies in their 'forgetfulness'. This feature seemingly makes them inappropriate for the modelling under consideration here. They overcome this limitation by incorporating memory into the state representation itself. In particular they describe what they term a history vector which is a representation of a patient's diagnostic history in the form of a binary vector which encodes the types of diagnoses that the patient has been given in the past.

3.2.1 Identifying confounding factors

Consider two history vectors, H_x and H_y, which differ in the presence of only a single past diagnosis d_d. In other words, all bits in H_x and H_y are the same except for exactly one. A specific follow-up diagnosis d_f, causes the transition of H_x and H_y to respectively H'_x and H'_y. We show how it can be automatically inferred if the differential diagnosis between h_x and h_y is one which affects the probability of d_f. We achieve this using a Bayesian approach which readily lends itself to asymmetrical risk driven inference, as described next. If the probability of d_f is not affected by the presence of d_d (in the context of other historical diagnoses in H_x and H_y, of course) then the transition data from the database of EHRs can be merged and thus used to estimate the aforesaid probability with higher precision so clearly this is a highly desirable goal which can be used to reduce the amount of confounding factors greatly and improve the accuracy of the learnt models.

Consider what happens if H_x and H_y are indeed merged in the context of the prediction of d_f. In such a case, the number of the observed transitions from H_x to $H_x \rightarrow d_f$ and from H_y to $H_y \rightarrow d_f$ are considered as equivalent. By considering them jointly a new probability of d_f from *either* H_x or H_y can be estimated. Call this probability z. The total risk ρ of the aforesaid merge can then be computed as a sum of risks associated with the actual probabilities of d_f following H_x and H_y respectively:

$$\rho = \rho_x + \rho_y. \tag{8}$$

This risk emerges as a consequence of the fact that the empirical nature of EHRs inherently involves a degree of stochasticity which means that there can never be absolute certainty that d_d is indeed entirely inconsequential in the context of this prediction. Instead, employing Bayesian framework, it is necessary to integrate over the latent probability of d_f following H_x and H_y and weight this with the associated relative risk. In this manner for ρ_x the risk can be written as:

$$\rho_x = C_x \int_z^1 |x - z| p(x|n_x) dx + \tag{9}$$

$$+ (1 - C_x) \int_0^z |z - x| p(x|n_x) dx. \tag{10}$$

What this expression captures can be readily understood as follows. The first term quantifies the risk of z *underestimating* the true probability x of d_f following H_x (hence the integration is for $x > z$). Similarly the second term quantifies the risk of z *overestimating* the true probability x of d_f following H_x (hence the integration is for $x < z$). The two risks are in general weighted asymmetrically, as governed by the constant $C_x \in [0, 1]$ which should be set by a relevant medical professional. The aforesaid asymmetry captures what are in general different 'costs' of overestimating and underestimating the probability of a particular diagnosis. For example, the cost of underestimating the probability of a terminal diagnosis is much greater than of overestimating it by the same amount. In this case C_x should be large i.e. closer to 1.

Continuing from (9), using Bayes theorem the term $p(x|n_x)$ can be rewritten as follows:

$$p(x|n_x) = \frac{p(n_x|x)p(x)}{p(n_x)}, \tag{11}$$

where n_x is the number of cases in which d_f was the next diagnosis following H_x, of the total of N_x transitions present in the EHRs database. Since the method has no means of establishing an informative prior on the transition probability x, an uninformative prior $p(x)$ is used which leads to $p(x) = 1$ since $x \in [0, 1]$. Moreover, $p(n_x|x)$ is readily identifiable as a binomial distribution with the parameter x and the number of draws N_x allowing $p(x|n_x)$ to be expanded further as follows:

$$p(x|n_x) = \frac{p(n_x|x)}{p(n_x)} \tag{12}$$

$$= \frac{\binom{N_x}{n_x} x^{n_x}(1-x)^{N_x-n_x}}{\int_0^1 p(n_x|w)dw} \tag{13}$$

$$= \frac{x^{n_x}(1-x)^{N_x-n_x}}{\int_0^1 \binom{N_x}{n_x} w^{n_x}(1-w)^{N_x-n_x} dw} \tag{14}$$

$$= \frac{x^{n_x}(1-x)^{N_x-n_x}}{\binom{N_x}{n_x} \beta(n_x+1, N_x-n_x+1)} \tag{15}$$

where $\beta(.)$ is the Euler beta function, and simple marginalization over x is performed in the denominator. This expression can be substituted back into (9) and (10), and then (8), and the integration performed numerically (which is both simple and fast, given that it is a simple integration in 1D).

Notes and remarks on practical application It is insightful to highlight several important practical aspects of the proposed technique. Firstly, once implemented as software it is intuitive to use – the tradeoff between over- and under-diagnosis is a concept routinely dealt with by medical professionals, and it is simply set using a single constant which balances the two risks. The risk is also readily interpretable. For example, for a terminal diagnosis the integrand

in (9) can be interpreted as computing the number of individuals who would be incorrectly expected to have a terminal diagnosis – an undesirable mistake considering the potential emotional stress, to begin with. Similarly, for a terminal diagnosis the integrand in (10) estimates the number of individuals who would experience a terminal episode which would not be predicted – arguably an even more serious mistake in that it *ipso facto* involves the loss of life. The acceptable tradeoff can be made by a clinician either on the level of an individual patient, for a specific diagnosis, or for an entire class of diagnoses (e.g. the same baseline risk tradeoff could be set for an entire ICD chapter, such as chapter IX which covers circulatory system diseases). In summary, the proposed technique is simple and intuitive to use, and it allows a high degree of flexibility in the choice of specificity or generality in application.

4 Evaluation

In this section we summarize some of the experiments we conducted to evaluate the proposed framework, and derive useful insights which illuminate possible avenues for improvement and future work.

4.1 EHR data

In an effort to reduce the possibility of introducing variability due to confounding variables, we sought to standardize our evaluation protocol as much as possible with that adopted by previous work. Hence we requested access to the large collection of EHRs described by (Arandjelović, 2015b) and were kindly pro-

vided 75% of the records used in the aforementioned paper. For completeness here we summarize the key features of this subset.

The EHRs adopted for evaluation were collected by a large private hospital in Fife, Scotland. The distribution of patient age in the database is 75 ± 14 years, the youngest and oldest patients being 18 months and 105 years old respectively, with the male to female ratio $56 : 44$. Approximately 23% of the patients in the database have a date of death associated with their EHR, which means that they are deceased and thus have a record of a terminal diagnosis. The entire EHR collection spans a period of 10 years, with the average number of diagnoses per patient of 9.9 ± 64.0.

4.2 Baseline model validation

Interestingly, on our data set the patient's age was found not to be associated with the number of admissions on record, while a low positive correlation ($r = 0.14$) was found between the patient's age and the number of conditions the patient had been diagnosed with at some point in the past – see Figures 4(b) and 4(b). A better predictor of the number of admissions was found to be the presence of a particular diagnosis (e.g. a high number of admissions is associated with the presence of the diagnoses of mental disorders, renal and cardiovascular conditions), as illustrated in Figures 5(a) and 5(b). Further insight can be gained by examining Figures 6(a) and 6(b) which summarize the repeated diagnosis statistics across different conditions. A mental disorder diagnosis or dialysis treatment for example predict both a high probability of a repeated di-

agnosis, as well as a high total number of the diagnosis type on record. These results are consistent with previous studies in the literature (Vigod et al., 2013; Kilkenny et al., 2013; Allaudeen et al., 2011) and support our diagnosis presence based model.

4.2.1 Next diagnosis prediction

To evaluate the predictive power of the proposed model, we examined its performance in the prediction of the next diagnosis based on a patient's prior diagnosis history, and compared this with the performance of the Markov process based approach described previously; see (1)–(3). Both methods were trained using an 80-20 split of data into training and test. Specifically, 80% of the data corpus was used to learn the model parameters – conditional probabilities $p(\hat{H} \rightarrow d|\hat{H})$ in the case of the proposed model and $p(d \rightarrow d')$ for the Markov process based model. The remaining 20% of the data was used as test input. For each test patient we considered the predictions obtained by the two methods given all possible partial histories. In other words, given a patient with the full diagnosis history $H = d_1 \rightarrow d_2 \rightarrow \ldots \rightarrow d_n$ we obtain predictions using partial histories $H_k = d_1 \rightarrow \ldots \rightarrow d_k$ for $k = 1 \ldots n-1$.

A summary of the results is given in Figure 7 which shows the cumulative match characteristic curves corresponding to the two methods – each point on a curve represents the proportion of cases (ordinate) for which the actual correct diagnosis type is at worst predicted with a specific rank (abscissa). The first thing that is readily observed from the plot is that the proposed method (blue line) vastly outperforms the Markov process based approach (red line). What is

more, the accuracy of our method is rather remarkable – it correctly predicts the type of the next diagnosis for a patient in 82% of the cases (rank-1). Already at rank-2 the accuracy is nearly 90%. In comparison, the Markov process based method achieves only 35% accuracy at rank-1, less than 50% at rank-2, and reaches 90% only at rank-17.

It is interesting to observe a particular feature of the CMC plot for the proposed method. Notice its tail behaviour – at rank-25 and above, the Markov process based approach catches up and actually performs better. While performance at such a high rank is not of direct practical interest, it is insightful to consider how this observation can be explained given that it is highly unlikely for it to be a mere statistical anomaly, considering the amount of data used to estimate the characteristics. The answer is readily revealed by considering the plot in Figure 8 which shows the dependency between the average rank of the proposed method's prediction and the length of the partial history used as input. Specifically, notice that higher ranks (i.e. worse performance) are associated with short histories. Put differently, when there is little information in a patient's history, there is more uncertainty about the patient's possible future ailments. This observation too strongly supports the validity of our model as it shows that accumulating evidence is used and represented in a more meaningful and robust way which allows for the learning of complex interactions between conditions and their development. Finally, this is illustrated in Figure 7 which also shows the plot of the proposed method's CMC curve restricted to test histories containing at least 5 prior diagnoses. In this case, rank-1 and rank-2 performances reach the remarkable accuracy of 91% and 97% respectively.

4.2.2 Long-term prediction

Given the outstanding performance of our method in predicting the type of the next diagnosis given the patient's current medical history, we next considered how the proposed model performs in long-term predictions. Considering that we are now dealing with sequences of future diagnoses and thus a much greater space of possible options, the characterization of performance using CMC curves is impractical. Rather, we now compare our approach with the Markov process based method by comparing the corresponding conditional probabilities for the actual progression observed in the data. In other words, for the prediction following a partial history $\hat{H}$ of the length k and the correct full history $H = \hat{H} \rightarrow d_{k+1} \rightarrow \ldots \rightarrow d_n$ we compute the log-ratio of conditional probabilities:

$$\rho = \log \left(\frac{p_{\text{Markov}}(\hat{H} \rightarrow d_{k+1} \rightarrow \ldots \rightarrow d_n | \hat{H})}{p_{\text{proposed}}(\hat{H} \rightarrow d_{k+1} \rightarrow \ldots \rightarrow d_n | \hat{H})} \right) \tag{16}$$

A positive value of ρ means that the Markov process based method performed better and a negative value that the proposed method did. The greater the absolute value of ρ the greater is the measured difference in performance in the corresponding direction. As before we divide the data into training and test sets using an 80-20 split and consider the predictions for all possible partial histories in the test set.

A summary of the results is presented in Figure 9. Specifically, the plot shows the cumulative distribution function (CDF) of the log-ratio ρ. As in the case of the one-step prediction, it is readily apparent that the performance of the proposed method vastly exceeds that of the Markov process based approach.

The value of CDF at the crossing of the curve with the $\rho = 0$ line is 0.82 which means that our method exhibited superior performance in 82% of the predictions. Even in the case of 18% of the predictions in which the Markov process based method performed better, the performance differential is not substantial. This is in sharp contrast with the instances in which the proposed method was better – in 67% of the cases the conditional probability of the correct history progression was over 100 greater for our model.

4.3 Assessing model scalability

Our primary goal here is to examine how the predictive performance of the history vector based model is affected by the choice of the number of salient diagnostic codes (Vasiljeva and Arandjelović, 2016a). As in (Arandjelović, 2015b) we too assess the quality of a specific prediction by considering the rank of the ground truth diagnostic code in the probability ordered list of predictions. Formally, let d_t be the ground truth diagnostic code which follows a particular history H. Then the rank r of d_t is given by the number of diagnostic codes which the model predicts as following H with at least the probability $p(H \rightarrow d_t)$:

$$r = |\{d : d \in D \wedge p(H \rightarrow d) \geq p(H \rightarrow d_t)\}|. \quad (17)$$

We used the same granularity of codes the original work described in (Arandjelović, 2015b).

Furthermore, we adopt the usual 'leave one out' evaluation protocol whereby the performance of the method is tested with each patient's data in turn and the model trained using the data of all other patients. To quantify the aggregate

performance of the model for specific model parameter values (i.e. the number of salient diagnoses included in the history vector representation) we use two well known measures. These are the average rank (a special case of the average normalized rank (Salton and McGill, 1983) when the set of target matches is exactly equal to 1) and the normalized area under the cumulative match characteristic (CMC) curve. For each possible rank r ($r = 1 \ldots n$, where n is the worst possible rank, equal to the number of diagnosis types), the CMC takes on the value equal to the proportion of predictions which predict the correct diagnosis at worst with the rank r (Bolle et al., 2005). The ideal performance results in the CMC having the value 1 across all ranks i.e. in each individual case the correct diagnosis is ranked 1. The area under the curve is normalized so that it is equal to 1 in this ideal case.

We started by looking at the effect that changing the number of salient diagnosis types, i.e. diagnosis codes with the corresponding (1-to-1) elements in the history vector, has on the area under the CMC curve. Our experimental results are captured by the plot in Figure 10(a). The plot can be readily seen to support our hypothesis that predicted a decay in the adopted model's prediction performance for an increasing number of explicitly modelled diagnoses. Notwithstanding this unwelcome qualitative observation, the major result is of a quantitative nature – the rate of the aforementioned decay is very slow indeed. Like many other natural phenomena the decay exhibits a power-law form with the associated exponent value which differs from 1 by only 5 parts in 100,000 i.e. it is equal to $1 - 0.5 \times 10^{-5}$. The practical significance of this finding is better appreciated by considering the plot in Figure 10(b). This plot shows the

variation in the area under the CMC curve as a function of the coverage of the entire diagnosis data corpus by the salient codes. The outstanding performance of the adopted method is illustrated well by noting, for example, that the dimensionality of history vectors can be increased to explicitly model the number of most frequent diagnosis codes which cover over 91% of the data, with the predictive performance of the method dropping by a mere 0.5% as compared to the coverage of only 61%. Even 98% of data coverage results in a change of only 0.8%. Recall that in the original paper the authors used 30 codes which accounted for 75% of the diagnoses in the corpus. Our results demonstrate that this was an overly conservative value.

We next examined the average prediction rank of the correct diagnosis type, which offers further insight into the performance of the adopted method. As expected from the previous set of findings, the results summarized by the plots in Figures 11(a) and 11(b) corroborate the observation that an increase in the dimensionality of history vectors, a key parameter of the method, worsens performance. In this experiment this worsening is exhibited as an increase in the average rank (i.e. a greater number of incorrect predictions are made with a higher probability than the actual ground truth diagnosis type). It is interesting to note the significance of what appears to be a much more rapid performance deterioration in terms of this performance measure in comparison with the area under the CMC curve discussed previously. For example, while the use of 200 vs. 10 most frequent diagnosis codes effects a reduction of only 0.5% in the area under the CMC curve, the corresponding change in the average rank of the correct diagnosis type increases fivefold (from approximately 1.5 for 10 salient

codes, to approximately 7.3 for 200 salient codes). The explanation for this apparent discrepancy is in fact reassuring as it demonstrates that the most dramatic changes in the predicted rank happen for predictions which are already not very good i.e. the small number of bad predictions become even worse, rather than good predictions becoming bad.

Lastly, to examine in additional detail how an increase in the number of explicitly modelled diagnosis types affects predictions, we looked at prediction rank histograms for different diagnosis codes and the corresponding changes as their number was changed. Figures 12(a) and 12(b) contrast the histograms for 20 and 50 salient diagnosis types. It is remarkable to observe that in both cases the histograms are virtually identical across different codes within the same model. Rather than being effected by sub-par histograms of the added codes, the (small, as demonstrated previously) deterioration in predictive performance as the number of salient diagnosis types is increased, is effected by slightly worse predictive performance uniformly distributed across different diagnoses. This is highly preferable in practice as it implies that for a fixed model complexity predictive power remains the same regardless of the patient's ailment. Were it otherwise, the predictions would be more difficult to interpret and the model complexity more challenging to set appropriately as the model's predictive performance would exhibit dependence on the nature of the health problems affecting a specific patient.

4.3.1 Assessing the effects of incorporating explicit clinical knowledge

Firstly we examined how the number of transition merges changes with the variation in the values of the two free parameters, namely the merging threshold t_m and the relative risk weighting constant C_x in (9) and (10). We applied our method to the entire EHRs data set though, as noted in the previous section, in practice it is likely that different parameters would be applied to different sub-trees of the diagnosis coding hierarchy.

Our findings are summarized by the surface plot shown in Figure 13. While it is inherently the case that increasing t_m cannot reduce the number of merges made, the characteristics of the corresponding change are insightful to the clinician in that they can be used to guide the choice of the risk weighting constant. Notice, for example, that the number of effected merges increases approximately linearly across the entire range of t_m for C_x smaller than approximately 0.5 whereas for C_x greater than 0.5 there is a much more sudden increase.

Next we examined salient diagnoses d_f (see Section 3.2) associated with the greatest number of merges. We noticed that the diagnosis of stroke was one of the particularly represented diagnosis amongst these, across different values of t_m and C_x, so we examined the corresponding merging behaviour in more detail. Interpreted intuitively, this means that on average the diagnosis of stroke has the least effect on (from the set of salient diagnoses included in the history vector) the prognosis of other ailments. The family of curves for different values of C_x, showing the variation of the number of merges (as the proportion of all possible transitions pairs which could possibly be merged and associated with

transitions effected by the diagnosis of stroke) as a function of the merging threshold t_m is shown in Figure 14. It is insightful to observe that much like in Figure 13, an increase in C_x results in more merges for the same value of t_m. A careful consideration of characteristics such as this one is crucial in the practical deployment of the proposed method, and the choice of granularity (in the context of the diagnosis coding hierarchy) at which the method is applied and its parameters.

5 Summary and future work

In this paper we introduced a novel algorithm that uses machine learning on EHR collections for the discovery of longitudinal patterns in the diagnoses of diseases. The two key technical novelties are: (i) a novel learning paradigm which enables greater learning specificity, and (ii) a method for risk driven identification of confounding diagnoses. A series of experiments were presented to demonstrate the effectiveness of the proposed techniques. Novel insights resulting from our experimental findings were also discussed and highlighted.

As regards possible future work directions, a number of possibilities were proposed by the authors of the original history vector based approach that the present method was partly inspired by. While we agree with most of these in broad terms, our contributions, experiments, and results suggest what we believe to be more promising immediate alternatives. In particular while we agree with the authors of the original method that the presence of a particular episode of care is a predictive factor not much weaker than the exact number of episodes

(which would require a prohibitively large amount of training data to learn), we believe that history vector binarization is an overly harsh step for the reduction of the learning space. Following the spirit of the method introduced in the present paper we intend to explore the possibility of automatically detecting chronic types of episodes of care (such as dialysis, for example) and then using a binary representation for non-chronic, and a more graded representation for chronic conditions.

Author disclosure statement

The authors have no competing interests to declare.

References

N. Allaudeen, A. Vidyarthi, J. Maselli, and A. Auerbach. Redefining readmission risk factors for general medicine patients. *J Hosp Med*, 6(2):54–60, 2011.

V. Andrei and O. Arandjelović. Identification of promising research directions using machine learning aided medical literature analysis. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2471–2474, 2016.

O. Arandjelović. Contextually learnt detection of unusual motion-based behaviour in crowded public spaces. *In Proc. International Symposium on Computer and Information Sciences*, pages 403–410, 2011.

O. Arandjelović. Prediction of health outcomes using big (health) data. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2543–2546, 2015a.

O. Arandjelović. Modelling disease progression using electronic hospital records. *In Proc. IJCAI Workshop on Bioinformatics and Artificial Intelligence*, pages 10–16, 2015b.

O. Arandjelović. On the discovery of hospital admission patterns – a clarification. *Bioinformatics*, 32(13):2078, 2016.

L. Barracliffe, O. Arandjelović, and G. Humphris. Can machine learning predict healthcare professionals' responses to patient emotions? *In Proc. International Conference on Bioinformatics and Computational Biology*, 2017.

N. Bartolomeo, P. Trerotoli, A. Moretti, and G. Serio. A Markov model to evaluate hospital readmission. *BMC Med Res Methodol*, 8(1):23, 2008.

D. M. Berwick and A. D. Hackbarth. Eliminating waste in US health care. *JAMA*, 307(14):1513–1516, 2012.

A. Bessou, F. Guelfucci, S. Aballea, M. Toumi, and C. Poole. Comparison of comorbidity measures to predict economic outcomes in a large UK primary care database. *Value Health*, 18(7):A691, 2015.

A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Data-mining Twitter and the autism spectrum disorder: a pilot study. *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 349–356, 2014.

A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Hierarchical Dirichlet process for tracking complex topical structure evolution and its ap-

plication to autism research literature. *In Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1:550–562, 2015a.

A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Using Twitter to learn about the autism community. *Social Network Analysis and Mining*, 5(1):5–22, 2015b.

A. Beykikhoshk, D. Phung, O. Arandjelović, and S. Venkatesh. Analysing the history of autism spectrum disorder using topic models. *In Proc. IEEE International Conference on Data Science and Advanced Analytics*, pages 762–771, 2016.

P. Bhatnagar, K. Wickramasinghe, J. Williams, M. Rayner, and N. Townsend. The epidemiology of cardiovascular disease in the UK 2014. *Heart*, 101(15): 1182–1189, 2015.

G. S. Birkhead, M. Klompas, and N. R. Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual Review of Public Health*, 36:345–359, 2015.

R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the ROC curve and the CMC. *In Proc. IEEE Workshop on Automatic Identification Advanced Technologies*, pages 15–20, 2005.

J. Butler and A. Kalogeropoulos. Hospital strategies to reduce heart failure readmissions. *J Am Coll Cardiol*, 60(7):615–617, 2012.

C. Canavan, J. West, and T. Card. Calculating total health service utilisation and costs from routinely collected electronic health records using the example of patients with irritable bowel syndrome before and after their first gastroenterology appointment. *PharmacoEconomics*, 34(2):181–194, 2015.

B. Christensen and G. Ellingsen. Evaluating model-driven development for large-scale EHRs through the openEHR approach. *Int J Med Inform*, 89: 43–54, 2016.

P. M. Coloma, G. Trifiro, V. Patadia, and M. Sturkenboom. Postmarketing safety surveillance : where does signal detection using electronic healthcare records fit into the big picture? *Drug Safety*, 36(3):183–197, 2013.

A. G. Crawford, C. Cote, J. Couto, M. Daskiran, C. Gunnarsson, K. Haas, and et al. Comparison of GE Centricity electronic medical record database and National Ambulatory Medical Care Survey findings on the prevalence of major conditions in the United States. *Popul Health Manag*, 13(3):139–150, 2010.

A. De Gaetano, T. Hardy, B. Beck, E. Abu-Raddad, P. Palumbo, J. Bue-Valleskey, and N. Pørksen. Mathematical models of diabetes progression. *Am J Physiol Endocrinol Metab*, 295:E1462–E1479, 2008.

K. Dharmarajan, A. F. Hsieh, Z. Lin, H. Bueno, J. S. Ross, I. Horwitz, J. A. Barreto-Filho, N. Kim, S. M. Bernheim, L. G. Suter, E. E. Drye, and H. M. Krumholz. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*, 309(4): 355–363, 2013.

N. D. Duffy and J. F. S. Yau. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Stat Med*, 14(14):1531–1543, 1995.

K. Fan, A. E. Aiello, and K. A. Heller. Bayesian models for heterogeneous personalized health data. *J Mach Learn Res*, 2016.

F. Folino and C. Pizzuti. Combining Markov models and association analysis for disease prediction. *Information Technology in Bio- and Medical Informatics*, pages 39–52, 2011.

B. Friedman, H. J. Jiang, and A. Elixhauser. Costly hospital readmissions and complex chronic illness. *Inquiry*, 45(4):408–421, 2008–2009.

K. R. Gabriel and J. Neumann. A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, 88 (375):90–95, 1962.

C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society, Series D*, 52(2):193–209, 2003.

M. F. Kilkenny, M. Longworth, M. Pollack, C. Levi, and D. A. Cadilhac. Factors associated with 28-day hospital readmission after stroke in Australia. *Stroke*, 44(8):2260–2268, 2013.

R. Kukafka, J. S. Ancker, C. Chan, J. Chelico, S. Khan, S. Mortoti, and et al. Redesigning electronic health record systems to support public health. *Journal of Biomedical Informatics*, 40(4):398–409, 2007.

E. C. Lau, F. S. Mowat, M. A. Kelsh, J. C. Legg, N. M. Engel-Nitz, H. N. Watson, and et al. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clin Epidemiol*, 3:259–272, 2011.

K. Lee, M. Ho, J. Yang, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

M. Liu, E. R. McPeek Hinz, M. E. Matheny, J. C. Denny, J. S. Schildcrout, R. A. Miller, and et al. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *J Am Med Inform Assoc*, 20(3):420–426, 2013.

N. Menachemi and T. H. Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, 4:47–55, 2011.

A. M. Mudge, K. Kasper, A. Clair, H. Redfern, J. J. Bell, M. A. Barras, G. Dip, and N. A. Pachana. Recurrent readmissions in medical patients: a prospective study. *J Hosp Med*, 6(2):61–67, 2011.

C. J. L. Murray, A. D. Lopez, C. D. Mathers, and C. Stein. The global burden of disease 2000 project: aims, methods and data sources. *World Health Organization*, 2001.

P. M. Nadkarni. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *J Am Med Inform Assoc*, 17(6):671–674, 2010.

R. Osuala and O. Arandjelović. Visualization of patient specific disease risk. *In Proc. IEEE International Conference on Biomedical and Health Informatics*, 2017.

M. M. Paul, C. M. Greene, R. Newton-Dame, L. E. Thorpe, S. E. Perlman, K. H. McVeigh, and et al. The state of population health surveillance using electronic health records: a narrative review. *Population Health Management*, 18(3):209–216, 2015a.

S. K. Paul, K. Klein, D. Maggs, and J. Best. The association of the treatment with glucagon-like peptide-1 receptor agonist exenatide or insulin with cardiovascular outcomes in patients with type 2 diabetes: a retrospective obser-

vational study. *Cardiovasc Diabetol*, 14(1):1–9, 2015b.

S. K. Paul, K. Klein, B. L. Thorsted, M. L. Wolden, and K. Khunti. Delay in treatment intensification increases the risks of cardiovascular events in patients with type 2 diabetes. *Cardiovasc Diabetol*, 14:100, 2015c.

RGI-CGHR Collaborators. Report on the causes of death in India: 2001–2003. *Office of the Registrar General of India*, 2009.

G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw Hill, New York, 1983.

R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using hidden Markov models. *In Proc. IEEE International Conference on Engineering in Medicine and Biology Society*, pages 2845–2848, 2012.

B. Topp, K. Promislow, G. de Vries, R. M. Miura, and D. T. Finegood. A model of β-cell mass, insulin, and glucose kinetics: Pathways to diabetes. *J Theor Biol*, 206:605–619, 2000.

I. Vasiljeva and O. Arandjelović. Prediction of future hospital admissions – what is the tradeoff between specificity and accuracy? *In Proc. International Conference on Bioinformatics and Computational Biology*, pages 3–8, 2016a.

I. Vasiljeva and O. Arandjelović. Towards sophisticated learning from EHRs: increasing prediction specificity and accuracy using clinically meaningful risk criteria. *In Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2452–2455, 2016b.

I. Vasiljeva and O. Arandjelović. Automatic knowledge extraction from EHRs. *In Proc. International Joint Conference on Artificial Intelligence Workshop*

on Knowledge Discovery in Healthcare Data, 2016c.

S. N. Vigod, V. H. Taylor, K. Fung, and P. A. Kurdyak. Within-hospital readmission: an indicator of readmission after discharge from psychiatric hospitalization. *Can J Psychiatry*, 58(8):476–481, 2013.

X. Wang, D. Sontag, and F. Wang. Unsupervised learning of disease progression models. *In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 85–94, 2014.

B. Wettermark, H. Zoega, K. Furu, M. Korhonen, J. Hallas, M. Norgaard, and et al. The Nordic prescription databases as a resource for pharmacoepidemiological research – a literature review. *Pharmacoepidemiol Drug Saf*, 22(7): 691–699, 2013.

J. A. Whittaker and M. G. Thomason. A Markov chain model for statistical software testing. *IEEE Transactions on Software Engineering*, 20(10):812–824, 1994.

World Health Organization. *International statistical classification of diseases and related health problems.*, volume 1. World Health Organization, 2004.

L. Xu, D. Wen, X. Zhang, and J. Lei. Assessing and comparing the usability of Chinese EHRs used in two Peking University hospitals to EHRs used in the US: A method of RUA. *Int J Med Inform*, 89:32–42, 2016.

W. Ye, D. J. M. Isaman, and J. Barhak. Use of secondary data to estimate instantaneous model parameters of diabetic heart disease: Lemonade Method. *Information Fusion*, 13:137–145, 2012.

Figures

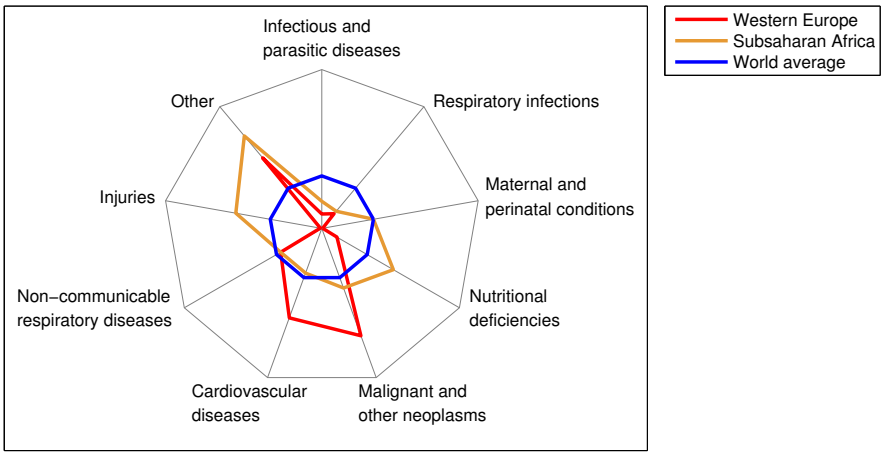


Fig. 1 Causes of death for the developed world (Western Europe), developing nations (Subsaharan Africa), and the world average.

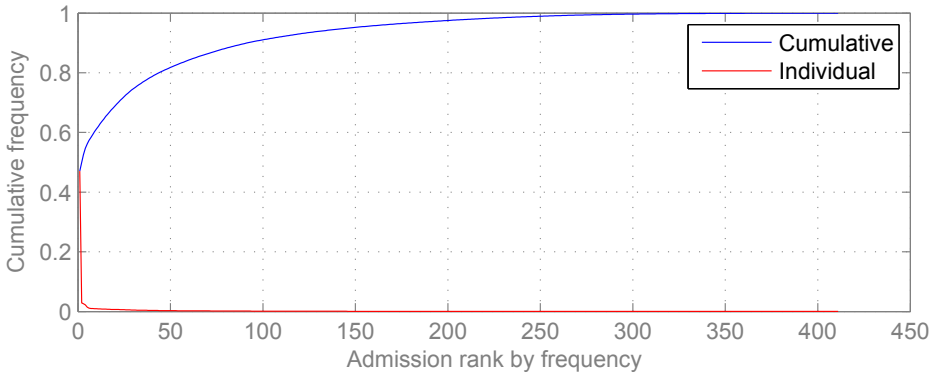


Fig. 2 Frequency (red line) and cumulative frequency of different diagnoses. The plot illustrates the highly uneven distribution, with the top 30 most frequent diagnoses accounting for 75% of the entire data corpus.

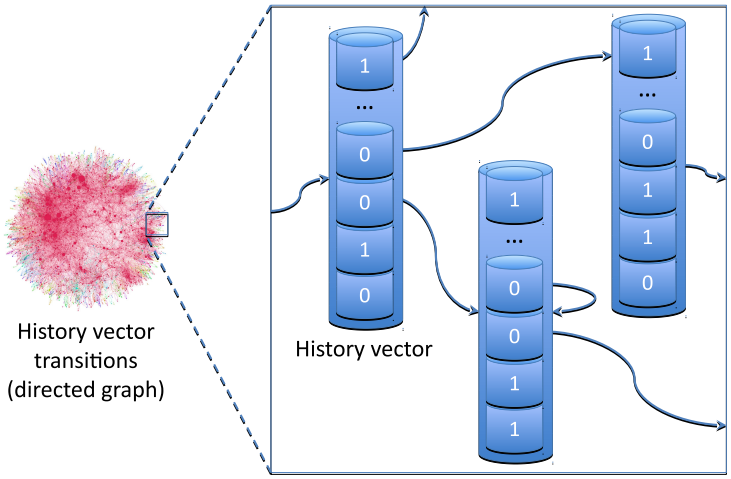
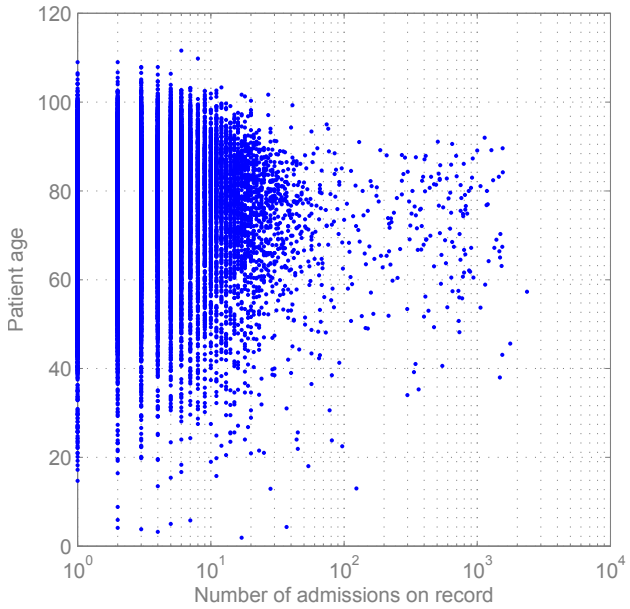
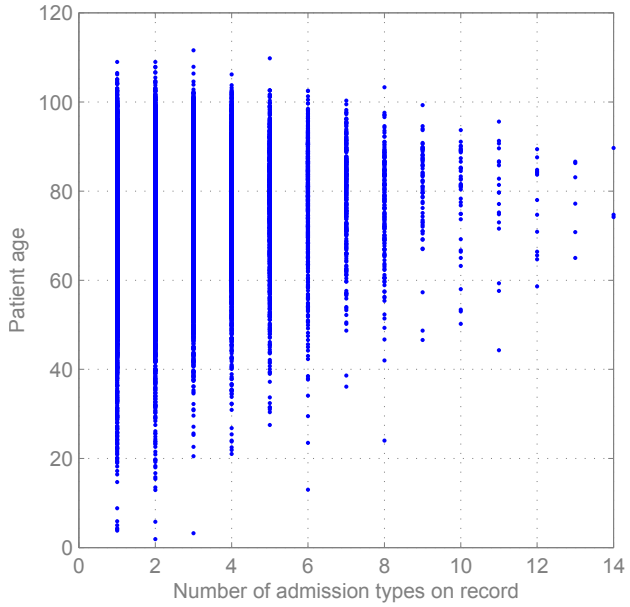


Fig. 3 Conceptual illustration of the method proposed by (Arandjelović, 2015b) which superimposes a Markovian model over a space of history vectors used to represent the medical state of a patient.

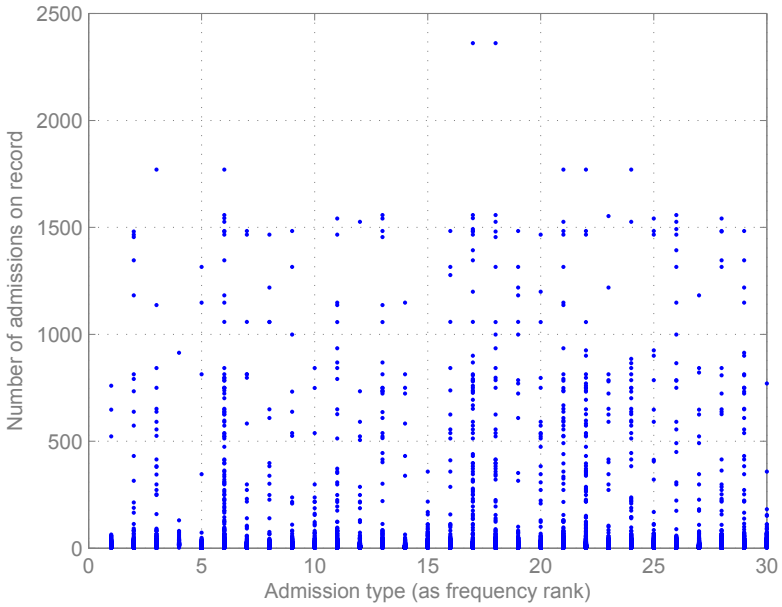


(a)

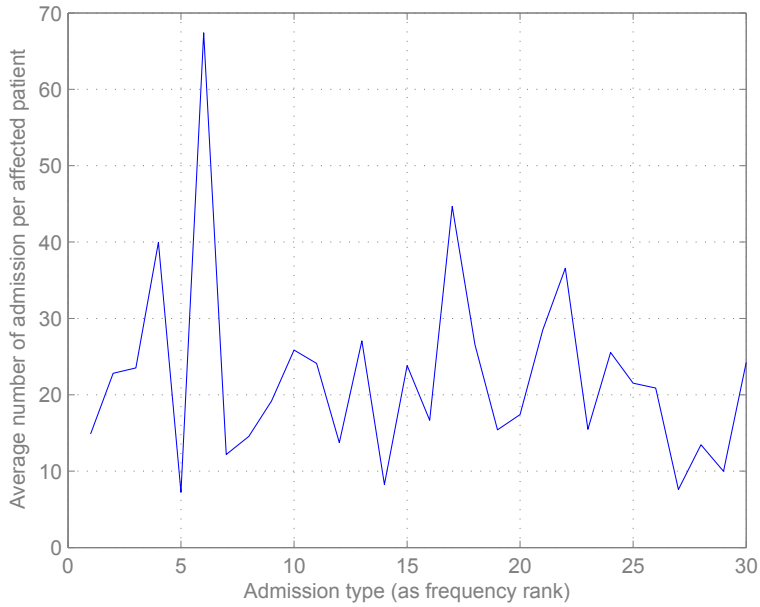


(b)

Fig. 4 (a) Patient age is not associated with the total number of admissions of the patient. (b) Patient age shows low association ($r = 0.14, p < 0.001$) with the number of conditions the patient has been diagnosed with.

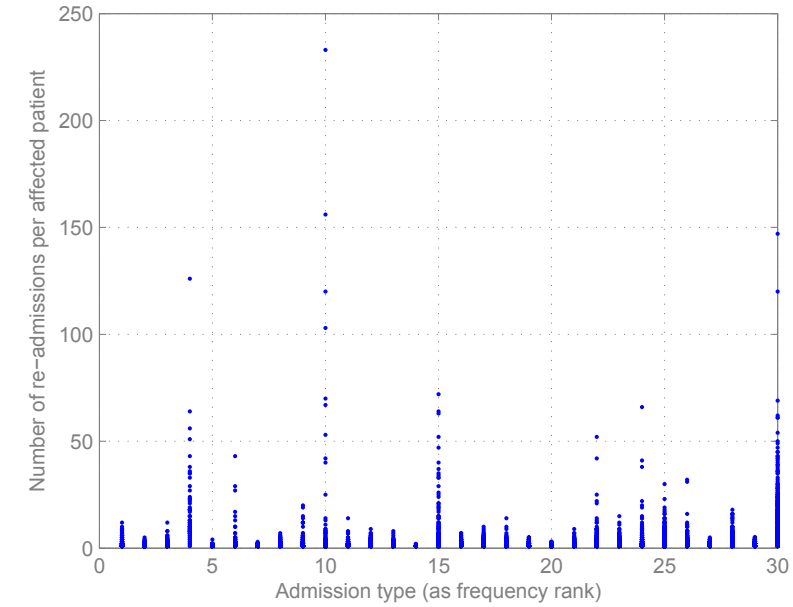


(a)

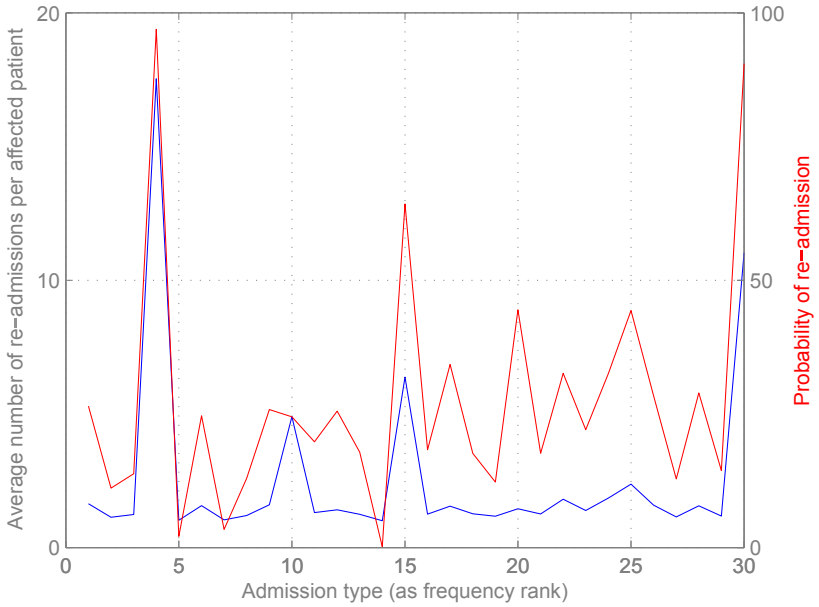


(b)

Fig. 5 (a) The presence of a particular condition in a patient's history is a good predictor of the total number admissions. (b) Average number of admissions for patients containing a particular diagnosed condition in their history.



(a)



(b)

Fig. 6 (a) Repeated diagnosis statistics for the top 30 diagnosed conditions. (b) Average number of repeated admussions and the probability of a repeated diagnosis for a particular condition.

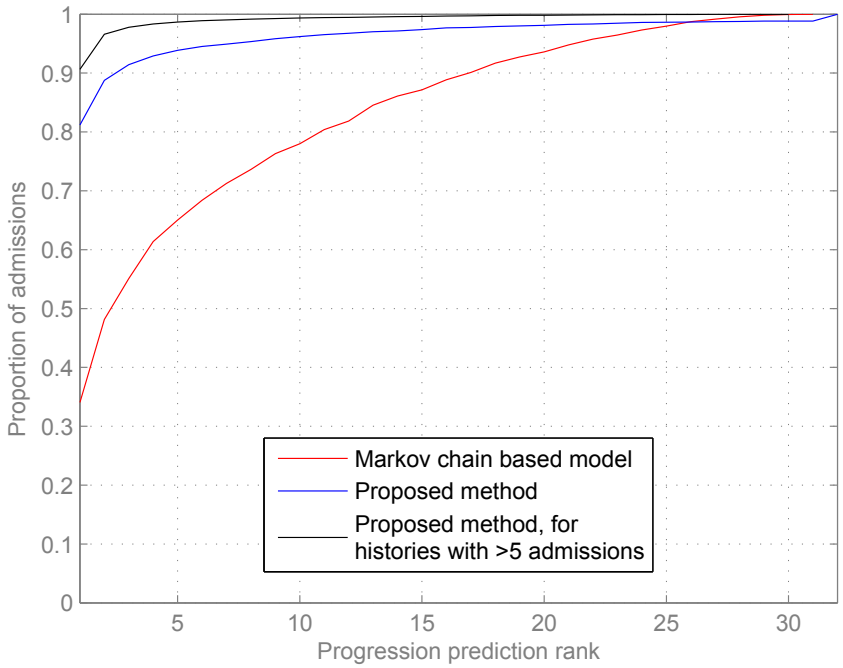


Fig. 7 Cumulative match characteristics (CMCs) for the prediction of the next diagnosis from a patient's history.

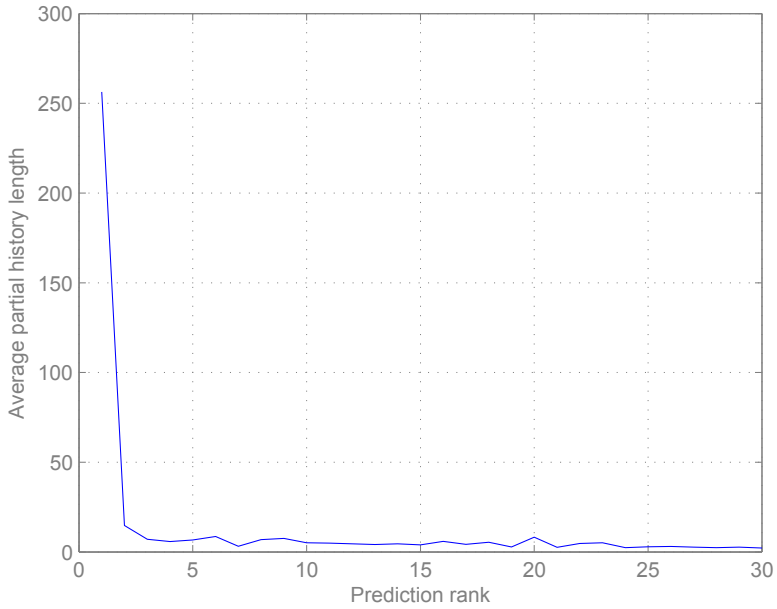


Fig. 8 Partial history length vs. next diagnosis prediction rank.

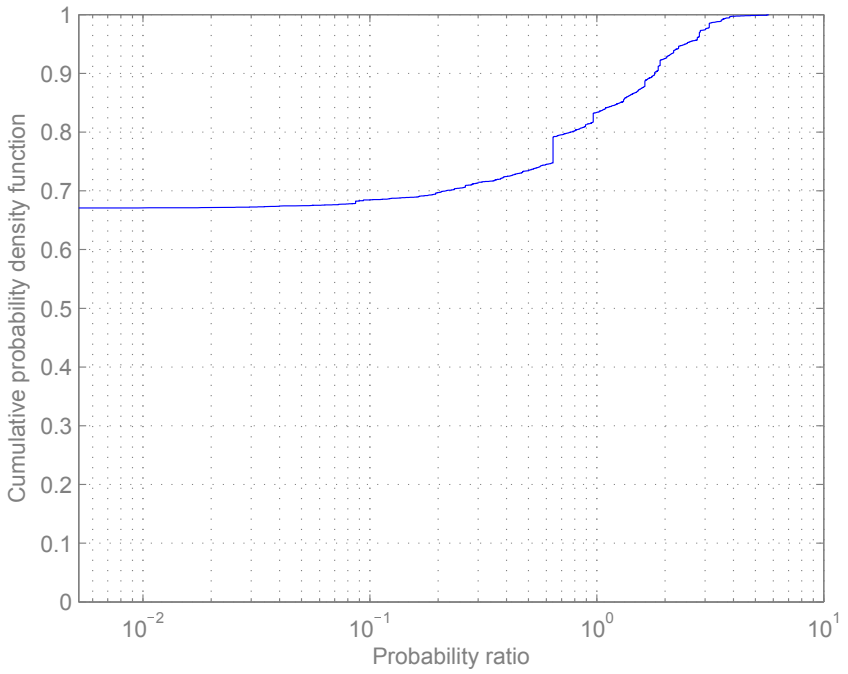
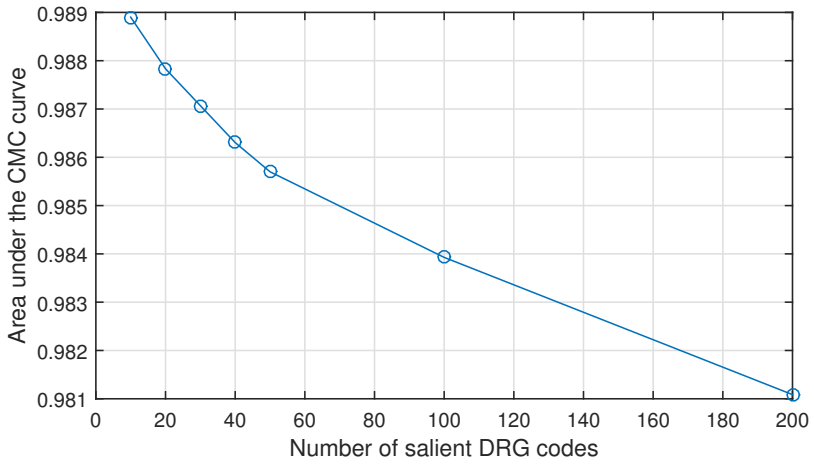
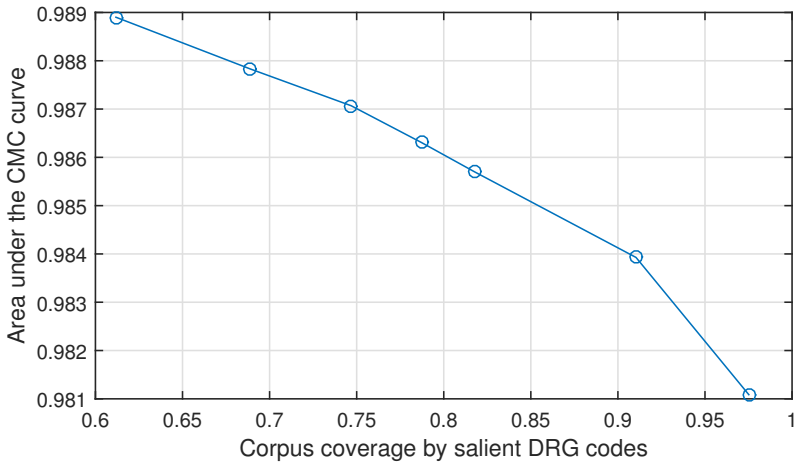


Fig. 9 Cumulative density function of the ratio of the probabilities of true patient medical history progression for the diagnoses-level Markov process approach and the proposed method.

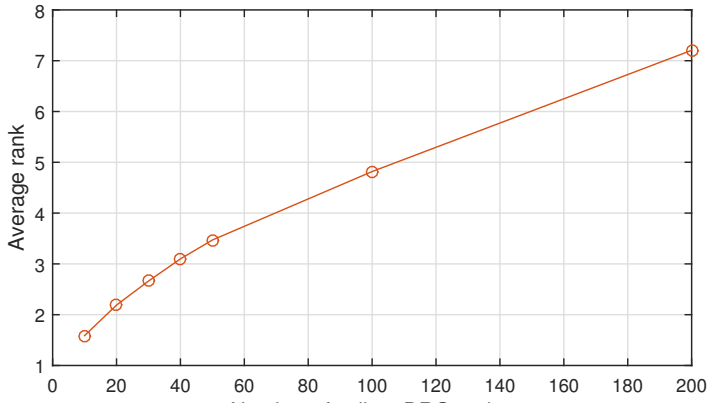


(a)

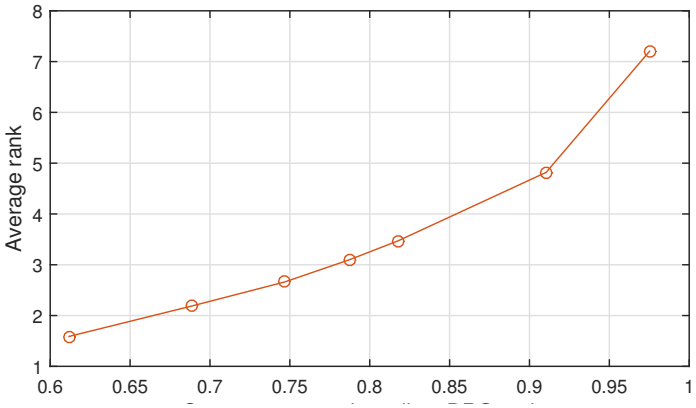


(b)

Fig. 10 The normalized area under the cumulative match characteristic (CMC) curve.

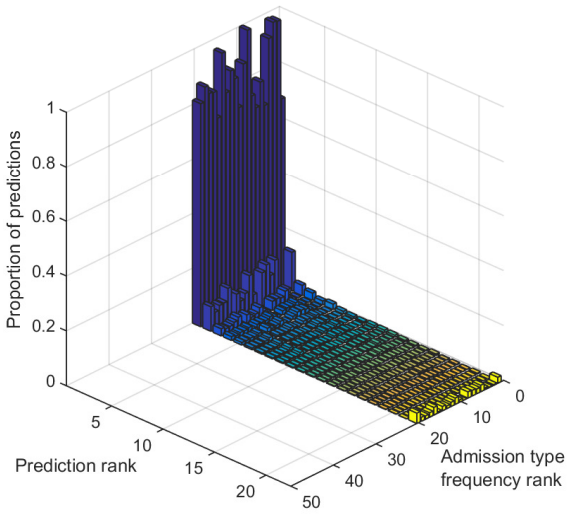


(a)

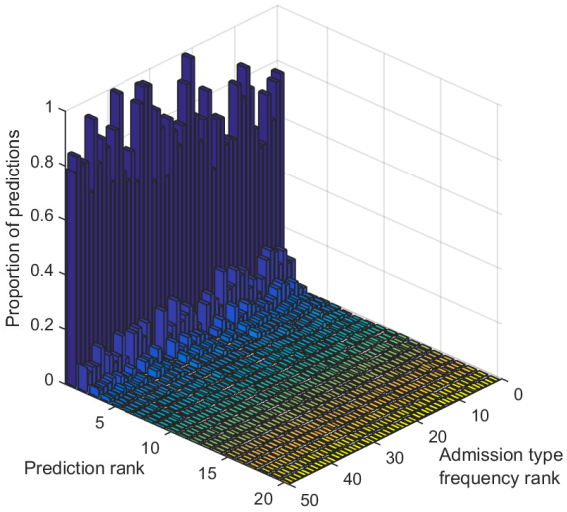


(b)

Fig. 11 The average prediction rank of the correct diagnosis type.



(a)



(b)

Fig. 12 Prediction rank histograms across different diagnosis codes using (a) 20 vs. (b) 50 salient codes.

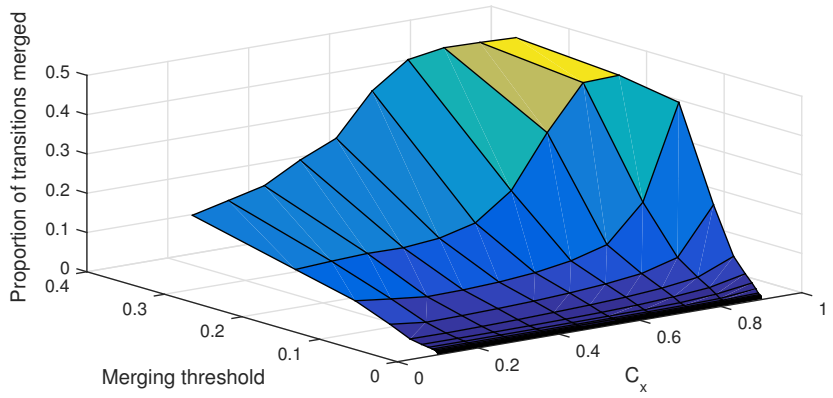


Fig. 13 Surface plot showing the number of pair-wise merges performed (as the proportion of all possible transitions pairs which could possibly be merged) as a function of the adjustable parameters of the proposed method, namely the merging threshold t_m and the relative risk weighting constant C_x in (9) and (10).

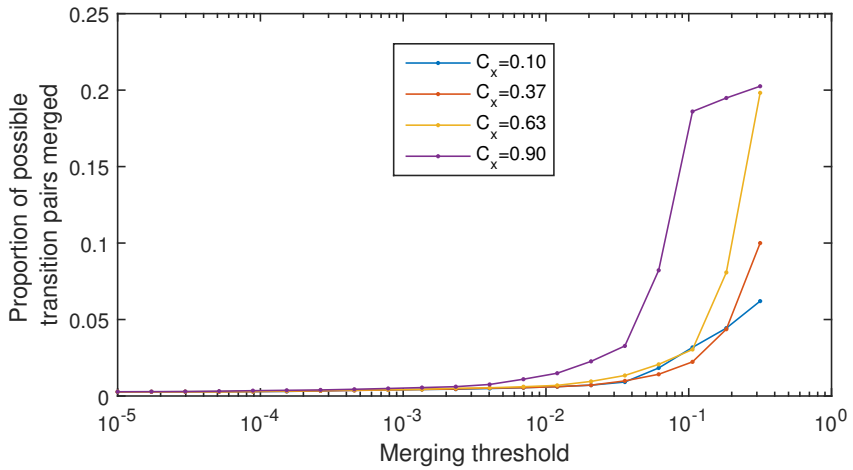


Fig. 14 The number of effected merges associated with the diagnosis of stroke (as d_f in Section 3.2) as the proportion of all possible transitions pairs which could possibly be merged and associated with transitions effected by the diagnosis of stroke.