Bioinformatics, 31(24), 2015, 3970–3976
doi: 10.1093/bioinformatics/btv508
Advance Access Publication Date: 3 September 2015
Original Paper

OXFORD

Data and text mining

Discovering hospital admission patterns using models learnt from electronic hospital records

Ognjen Arandjelović

School of Computer Science, University of St Andrews, St Andrews, UK

Associate Editor: Jonathan Wren

Received on May 6, 2015; revised on August 22, 2015; accepted on August 24, 2015

Abstract

Motivation: Electronic medical records, nowadays routinely collected in many developed countries, open a new avenue for medical knowledge acquisition. In this article, this vast amount of information is used to develop a novel model for hospital admission type prediction.

Results: I introduce a novel model for hospital admission-type prediction based on the representation of a patient's medical history in the form of a *binary history vector*. This representation is motivated using empirical evidence from previous work and validated using a large data corpus of medical records from a local hospital. The proposed model allows exploration, visualization and patient-specific prognosis making in an intuitive and readily understood manner. Its power is demonstrated using a large, real-world data corpus collected by a local hospital on which it is shown to outperform previous state-of-the-art in the literature, achieving over 82% accuracy in the prediction of the first future diagnosis. The model was vastly superior for long-term prognosis as well, outperforming previous work in 82% of the cases, while producing comparable performance in the remaining 18% of the cases.

Availability and implementation: Full Matlab source code is freely available for download at: http://ognjen-arandjelovic.t15.org/data/dprog.zip.

Contact: ognjen.arandjelovic@gmail.com

1 Introduction

Public healthcare is an issue of major global significance and concern. On the 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 toward so-called 'diseases of affluence'. Healthcare 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, the medical research community, healthcare practitioners and patients (Axon and Williams, 2011).

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 modern diseases of the developed world. The key premise of this 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 healthcare services in industrially and technologically developed countries. Herein, the specific aim is to develop a framework which allows a health practitioner (e.g. a doctor or a clinician) to understand the available patient information in an intuitive yet powerful fashion. Such a framework would, on the 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 healthcare providers in most developed countries. Specifically, this article used medical records of over 40 000 individuals collected by a local hospital (at present, this dataset cannot be made publicly available). This is a rich dataset, 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 this 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 nature and the format of this data are described 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 international 'codebook', the International Statistical Classification of Diseases and Related Health Problems (ICD) (World Health Organization, 2004), a medical classification list of diseases, injuries, symptoms, examinations, physical, mental or social circumstances issued by the World Health Organization. The ICD has a tree-like structure; at the top-most level codes are grouped into 12 chapters, each chapter encompassing a spectrum of related health issues (usually symptomatically rather than etiologically related). For example, 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 article, the classification attained at the depth of two is used, which achieves a good compromise between specificity (and thus medical significance and interest) and frequency of occurrence (thus ensuring that sufficient data are available that meaningful patterns can be discovered). This results in each admission being given a three character code which comprises a leading capital letter (A–Z, first classification level), followed by a two digit number (further refinement). For example, E62 codes for 'Respiratory infections/inflammations' within the broader range of conditions falling under the umbrella of 'Endocrine, nutritional and metabolic diseases'.

2 Sequential modeling

The major contribution of this work is a novel model for the prediction of future hospital admissions. 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 (Arandjelović, 2015).

Considering the importance of the problem at hand, it is unsurprising that it has attracted a significant amount of research attention. Most existing methods constrain their prediction to a narrow specific context, e.g. to admissions to the emergency department (Li and Guo, 2009), to heart failure-related admissions (Hammill *et al.*, 2011) and to the veteran population (Holloway *et al.*, 1990). The applicability of these methods is further limited by their frequent reliance on a substantial amount of expert knowledge in the choice of variables used for prediction (Leegon *et al.*, 2005). Notwithstanding these efforts, the performance of the methods described in the literature has largely been disappointing (Kansagara *et al.*, 2011). Better results have been reported in prediction attempts which simplify the task even further by looking at short-term (usually of ∼30 days) predictions only (Holman *et al.*, 2005).

One of the possible reasons for the poor performance of the existing methods in the literature lies in their virtually universally overly simplistic inference models. In particular, unlike in this work, they fail to capture sequential information on historical admissions and diagnoses; rather, they base their predictions on a single cumulative snapshot of a patient's record (Bottle *et al.*, 2006).

2.1 Markovian models

Although this work addresses the problem of hospital admission prediction, the sequential modeling paradigm at the center of the proposed method shares some common features with a number of models aimed at capturing disease progression patterns. Being readily adaptable to the task at hand, these models present a sensible baseline against which the performance of proposed method can be gauged so they are briefly reviewed next.

In contrast to most disease progression models which focus on specific individual diseases, such as type II (adult-onset) diabetes mellitus (De Gaetano *et al.*, 2008; Topp *et al.*, 2000) or heart disease (Ye *et al.*, 2012), and which are inherently 'low-level'-based in the sense that they explicitly model known physiological changes that affect disease progression, the methods of interest here model disease progression as a discrete sequence of events, with the progression governed by what is assumed to be a first-order Markov process (Jackson *et al.*, 2003; Sukkar *et al.*, 2012). The same idea is easily applied to hospital admission modeling by considering admissions as events and the patient's admission history as the corresponding sequence $H = a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_n$ where a_i is a discrete variable whose value is an ICD code corresponding to the ith of n admissions on the patient's record (which I will henceforth refer to as the length of H). The parameters of the underlying first-order Markov model can then be learnt by estimating the transition probabilities $p(a' \rightarrow a'')$ 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) (Bartolomeo *et al.*, 2008; Folino and Pizzuti, 2011; Wang *et al.*, 2014). The model can be applied to predict the nature of the admission a_{n+1} expected to follow from the current history by likelihood maximization:

$$a_{n+1} = \arg\max_a p(a_n \rightarrow a). \tag{1}$$

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

$$p_f(a^*) = \sum_a [p(a \rightarrow a^*)\ p_f(a)], \tag{2}$$

or to sample the space of possible hospital admission histories:

$$H' = a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_n \dashrightarrow a_{n+1} \dashrightarrow a_{n+2} \ldots . \tag{3}$$

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 modeled phenomena include meteorological events (Gabriel and Neumann, 1962), software usage patterns (Whittaker and Thomason, 1994), breast cancer screening (Duffy and Yau, 1995), human behavior (Arandjelović, 2011) and many others. Nonetheless, the key premise motivating the model described in this study is that the Markovian assumption is in fact not appropriate for modeling

hospital admission patterns (note that I do not reject its possible applicability in disease progression modeling on different levels of abstraction). Indeed, I will demonstrate this empirically. My claim 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 admission lost. The model proposed in the section which follows solves this problem while at the same time retaining the tractability of Markov process-based approaches.

2.2 Proposed approach

Distilling down the problem at hand to its central challenge, we wish to predict the probability of a specific admission a directly given and following a specific 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 H is infinite as H can be of an arbitrary length. For example, in the dataset used in this article, 0.4% of the patients have had more than 1000 admissions each. Even if the length $l(H)$ of H is limited, the number of possible histories is extremely large; specifically, it is $[l(H)]^{n_a}$, where n_a is the number of different admission codes. Therefore, it is necessary to make an approximation which constrains and simplifies the task. I already argued why the Markovian assumption on the level of admission codes is inappropriate. Instead, I propose a different representation of a patient's state, particularly suitable for the modeling of hospital admission patterns as understood in the present context. Consider a particular admission history $H = a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_n$. My 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 (Butler and Kalogeropoulos, 2012; Dharmarajan *et al.*, 2013; Friedman *et al.*, 2008–2009; Mudge *et al.*, 2011). Thus, I represent a history H using a history vector $v = v(H)$ which is a *fixed-length* vector with *binary* values. Each vector element corresponds to a specific admission code (except for one special element whose purpose will be explained shortly) and its value is 1 if and only if the corresponding admission is present in the history and 0 otherwise:

$$\forall a \in A.\ v(H)_{i(a)} = \begin{cases} 1 : \exists j.\ H = H_1 \rightarrow a_j \rightarrow H_2 \wedge a = a_j \\ 0 : \text{otherwise} \end{cases}, \tag{5}$$

where A is the set of possible admission codes, $i(a)$ indexes the admission code a in a history vector and H_1 and H_2 may take on degenerate forms of empty histories (i.e. no recorded admissions).

It is worth highlighting a few key aspects of this representation. Consider the choice of using a fixed-length vector to represent a patient's state (observed as a history of admissions). By collapsing a history of an arbitrary length 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 admissions is lost. Next, consider the binary nature of the vector elements. This feature 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 re-admissions 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 related admissions.

The modeling problem is thus reduced to the task of learning transition probabilities between different patient history vectors:

$$p(v(H) \rightarrow v(H')). \tag{6}$$

It is important to observe that unlike in the case of Markov process models working on the admission level when the number of possible transition probabilities is close to $n_a{}^2$ (the few transitions which are not possible are those which defy some basic causality rules, e.g. that diagnosis of a condition always precedes its treatment), 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 admission to one which does not, that is:

$$v(H)_{i(a)} = 1 \wedge v(H')_{i(a)} = 0 \Rightarrow p(v(H) \rightarrow v(H')) = 0. \tag{7}$$

The converse does not hold, however. Possible transitions can be only those which include either no changes to the history vector (re-admission) or which encode at least one additional admission:

$$p(v(H) \rightarrow v(H')) = \begin{cases} \phi : \forall a.\ v(H)_{i(a)} = 1 \Rightarrow v(H')_{i(a)} = 1 \\ \quad \text{and} \\ \quad |\{a : v(H)_{i(a)} = 1\}| \leq |\{a : v(H')_{i(a)} = 1\}| \\ 0 : \text{otherwise} \end{cases} \tag{8}$$

where $\phi \geq 0$. 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 dataset used in this article), 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 (in this context) low sample size. For this reason, admission history vectors are constructed using only the $\hat{n}_a$ most common admission types and the remaining $n_a - \hat{n}_a$ types merged into a single special code 'other'. Thus, the dimensionality of an admission history vector becomes $\hat{n}_a + 1$. The soundness of this approach can be readily observed by examining the plot in Figure 1 which shows that only a small number of admission types explains a vast proportion of all data. For example, the top 30 most frequent types account for 75% of all admissions. These are listed in Table 1, ranked and with the corresponding ICD codes and verbal descriptions.

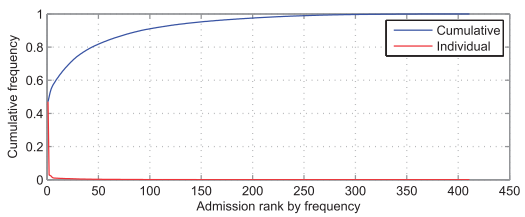

Fig. 1. Frequency (red line) and cumulative frequency of different admissions. The plot illustrates the highly uneven distribution, with the top 30 most frequent admissions accounting for 75% of the entire data corpus

Table 1. ICD codes of the 30 most frequent admission types in the dataset used in this study

Freq. rank	ICD code	Short description
1	R63	Chemotherapy
2	F42	Circulat disorders w/o AMI w invasive cardiac inves proc
3	F74	Chest pain
4	B70	Stroke
5	F62	Heart failure and shock
6	E65	Chronic obstructive airways disease
7	K60	Diabetes
8	F72	Unstable angina
9	Q61	Red blood cell disorders
10	G67	Oesophagitis, gastroent and misc digest system disorders
11	C16	Malignant neoplasm of the stomach
12	F66	Coronary atherosclerosis
13	E62	Respiratory infections/inflammations
14	F14	Vascular procs except major reconstruction
15	F73	Syncope and collapse
16	R61	Lymphoma and non-acute leukemia
17	F06	Coronary bypass w/o invasive cardiac inves proc
18	G44	Other colonoscopy
19	F71	Non-major arrhythmia
20	L63	Kidney and urinary tract infections
21	B71	Cranial and peripheral nerve disorders
22	I68	Non-surg neck and back cond w/o pain managmt proc/myelo
23	F60	Circulat disorders w AMI w/o invasive cardiac inves proc
24	F41	Circulat disorders w AMI w invasive cardiac inves proc
25	L67	Other kidney and urinary tract diagnoses
26	F10	Percutaneous coronary angioplasty w AMI
27	U40	Mental health treatment
28	G66	Abdominal pain or mesenteric adenitis
29	X60	Injury
30	Z40	Follow up after completed treatment w endoscopy

These account for 75% of all admissions.

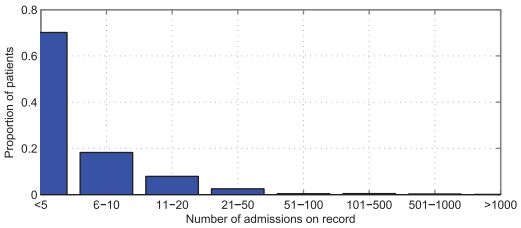

Fig. 2. Distribution of the number of admissions per patient across the dataset used in this article. Approximately 70% of the patients had up to 5 admissions, 88% up to 10, 96% up to 20 and 99% up to 50. A small but significant number of patients had a very large number of admissions on record, with 0.4% of patients with over 1000 admissions

2.3 Empirical model validation

Having described the model that is used to learn hospital admission patterns from electronic medical records, I now turn my attention to the validation of the model on real data and the comparison of its performance with that of Markov process-based methods previously proposed in the literature. The data used for this purpose are the set of medical records of over 40 000 people treated by a local hospital.

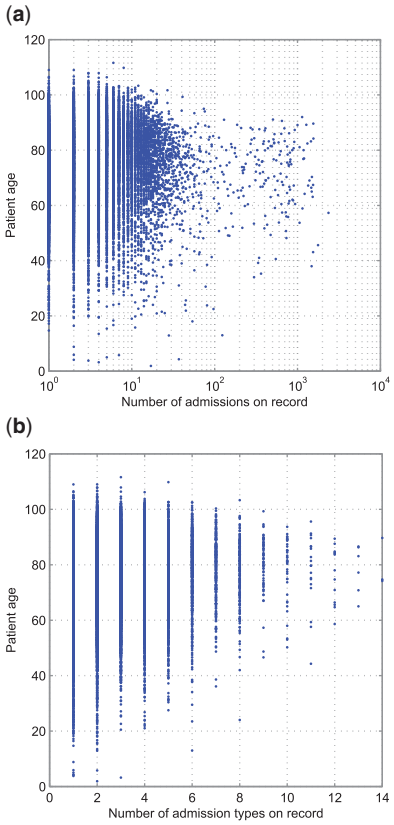

Fig. 3. (**a**) Patient age at the time of the last admission is not associated with the total number of past admissions of the patient. (**b**) Patient age shows low association ($r = 0.14, P < 0.001$) with the number of conditions the patient had been diagnosed with at some point in the past. In both plots, each point (i.e. respectively, patient age and number of admissions, or patient age and number of admission types pair) corresponds to a single of 40 000 individuals in the corpus of medical records used in this article

It is insightful to consider some of the characteristics of this dataset before proceeding with the analysis. As expected, the number of admissions per patient was found to vary greatly across the sample; the plot in Figure 2 shows the distribution of the data across different ranges of the number of admissions. Approximately 70% of the patients had up to 5 admissions, 88% up to 10, 96% up to 20 and 99% up to 50. A small but significant number of patients had a very large number of admissions on record, with 0.4% of patients with over 1000 admissions.

Interestingly, the patient's age was found not to be associated with the number of past 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 3b and 4a. A better predictor of the number of admissions was found to be the presence of a particular diagnosis/condition, as illustrated in Figure 4a and b. Observe, e.g. the high number of admissions associated with the presence of the diagnoses of mental disorders, renal and cardiovascular conditions. Further insight can be gained by examining Figure 5a and b, which summarizes the re-admission statistics across different conditions. A mental disorder diagnosis or dialysis treatment, e.g. predict both a high probability of re-admission and a high total number of re-admissions. These results are consistent with previous studies in the literature (Allaudeen *et al.*, 2011; Kilkenny *et al.*, 2013; Vigod *et al.*, 2013) and strongly support the diagnosis presence-based progression model proposed in this article.

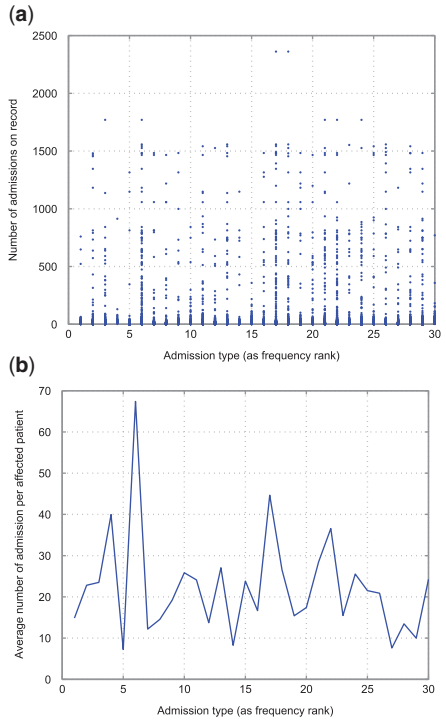

Fig. 4. (**a**) The presence of a particular condition in a patient's history is a good predictor of the total number of the patient's admissions. Each point (i.e. diagnosed condition and number of admissions pair) in the plot corresponds to a single of 40 000 individuals in the corpus of medical records used in this article. (**b**) Average number of admissions for patients containing a particular diagnosed condition in their history

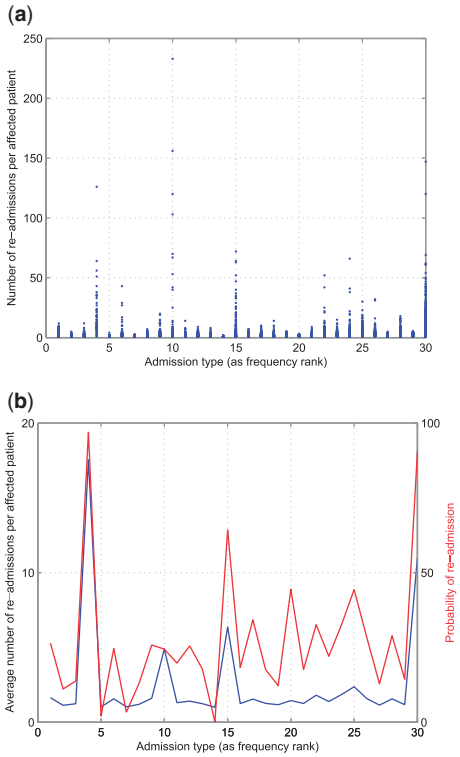

Fig. 5. Re-admission statistics for the top 30 diagnosed conditions (Table 1). (**a**) Each point (i.e. diagnosed condition and number of corresponding re-admissions pair) in the plot corresponds to a single of 40 000 individuals in the corpus of medical records used in this article. (**b**) Average number of re-admissions and the probability of re-admission for a particular condition for patients who were diagnosed with the condition at some point in the past

2.3.1 Next admission prediction

To evaluate the predictive power of the proposed model, I examined its performance to predict the type of the next admission based on the patient's prior admission history and compared this with the performance of the Markov process-based approach described previously; see (1)–(3). Both methods were trained using 100 instances involving random 80:20 splits of data into training and test corpora. Specifically, in each instance, 80% of the entire data corpus was used to learn the model parameters—the conditional probabilities $p(\dot{H} \rightarrow a|\dot{H})$ in the case of the proposed model and $p(a \rightarrow a')$ for the Markov process-based model. The remaining 20% of the data were used as novel, test input. For each test, patient I considered the predictions obtained by the two methods given all possible partial histories. In other words, given a patient with the full admission history $H = a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_n$, I obtain predictions using partial histories $H_k = a_1 \rightarrow \ldots \rightarrow a_k$ for $k = 1 \ldots n-1$, i.e. $H_1 = a_1$, $H_2 = a_1 \rightarrow a_2$, $H_3 = a_1 \rightarrow a_2 \rightarrow a_3$ etc.

A summary of the obtained results in given in Figure 6. The plot shows the cumulative match characteristic (CMC) curves corresponding to the two methods—each point on a curve represents the proportion of cases (the value of the ordinate) for which the actual correct admission type is at worst predicted with a specific rank (the value of the 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 the proposed method is rather remarkable—it correctly predicts the type of the next admission for a patient in 82% of the cases (rank 1 performance). Already at rank 2 (i.e. the correct admission type is one of the two predicted as the most probable), the accuracy is nearly 90%. In comparison, the Markov process-based

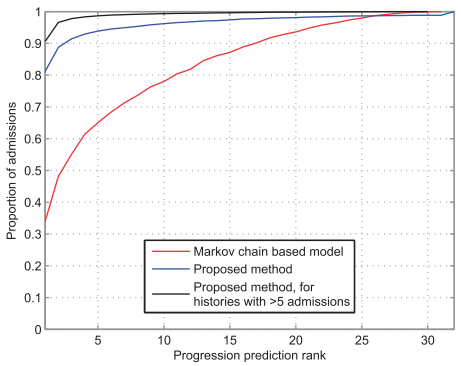

Fig. 6. The CMC curves for the prediction of the next admission-based on a patient's current history. The proposed model (blue line) vastly outperforms the Markov process-based method (red line). Black line shows the CMC curve for the proposed method when it is applied in the prediction using partial histories containing at least five prior admissions

method achieves only 35% accuracy at rank 1, less than 50% at rank 2, and reaches 90% only at rank 17. As expected, more primitive strategies perform worse: predicting that the next admission is the same as the admission which precedes it results in rank 1 success rate of 7.4%, random guessing without accounting for relative frequencies of different admissions in rank 1 success rate of 2.7% and random guessing while accounting for relative frequencies of different admissions in rank 1 success rate of 11.2%.

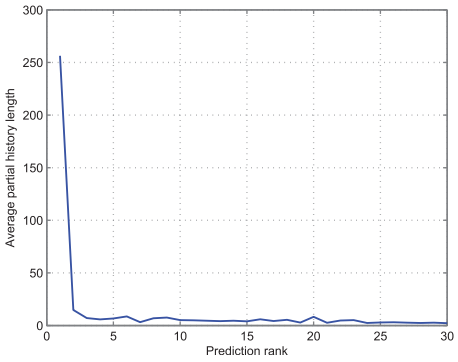

Fig. 7. The average partial history length as a function of the proposed model's next admission prediction rank. Notice that in all cases in which the proposed method did not correctly predict the next admission type (i.e. rank value is >1), the average partial history length is small, i.e. there are few prior admissions to base the prediction on. This observation strongly supports the validity of the proposed representation and model—it shows that accumulating evidence, when available, is used and represented in a more meaningful and robust way which allows for the learning of complex interactions between conditions and their development. When there is little information in a patient's history, there is understandably more uncertainty about the patient's possible future ailments

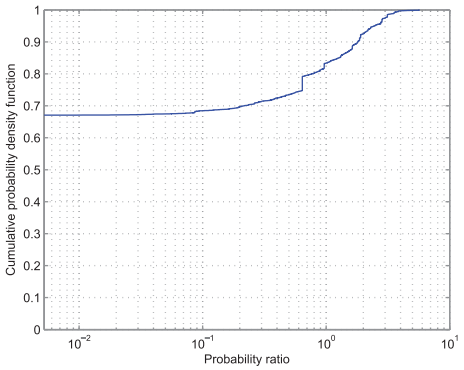

Fig. 8. Cumulative density function of the ratio of the probabilities of true patient medical history progression for the admission-level Markov process-based approach and the proposed method. The probabilities are computed for the second half of a patient's admissions record conditioned on the first half. Notice that the abscissa scale is logarithmic. As the plot illustrates, in 82% of the cases the proposed method exhibits superior performance. The vertical dotted red line (at the ratio $10^0 = 1$) marks the transition at which the proposed method ceases to perform better. It is important to observe that even in the 18% of the cases in which the Markov process-based approach yields better results, the proposed method is highly competitive. In contrast, in the majority of cases, the proposed method not only performs better but also with a high margin

It is interesting to observe a particular feature of the CMC plot for the proposed method. Notice its tail behavior—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 7 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 the proposed 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 6 which also shows the plot of the proposed method's CMC curve restricted to test histories containing at least five prior admissions. In this case, rank 1 and rank 2 performances reach the remarkable accuracy of 91% and 97%, respectively.

2.3.2 Long-term prediction

Given the outstanding performance of the proposed method in predicting the type of the next admission given the patient's current medical history, next I considered how the model performs at long-term predictions. Considering that we are now dealing with sequences of future admissions and thus a much greater space of possible options, the characterization of performance using CMC curves is impractical. Rather, I now compare my 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 $H = \hat{H} \rightarrow a_{k+1} \rightarrow \ldots \rightarrow a_n$ I compute the log-ratio of conditional probabilities:

$$\rho = \log\left(\frac{p_{\text{Markov}}(H|\hat{H})}{p_{\text{proposed}}(H|\hat{H})}\right) \tag{9}$$

$$= \log\left(\frac{p_{\text{Markov}}(\hat{H} \rightarrow a_{k+1} \rightarrow \ldots \rightarrow a_n|\hat{H})}{p_{\text{proposed}}(\hat{H} \rightarrow a_{k+1} \rightarrow \ldots \rightarrow a_n|\hat{H})}\right) \tag{10}$$

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 I performed 100 experiments, in each dividing the data into training and test sets using an 80:20 split and considering the predictions for all possible partial histories in the test set.

A summary of the results is presented in Figure 8. Specifically, the plot shows the cumulative distribution function of the log ratio ρ (note that the scale of the abscissa is logarithmic). 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. Notice that the value of cumulative distribution function at the crossing of the curve with the $\rho = 0$ line is 0.82 which means that the proposed method exhibited superior performance in 82% of the predictions. It is equally important to observe that 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 the proposed model.

3 Summary and conclusions

In this article, the goal was to develop a framework for the inference of complex hospital admission patterns from electronic medical

records of patients routinely collected by healthcare providers in the developed world. This wealth of information has tremendous potential in increasing the understanding of disease etiology, interactions of different risk factors and symptoms, etc.

I described two major contributions. The first of these is a novel model of hospital admission patterns. A crucial consideration in the development of this model was that, on the one hand, it must be sufficiently flexible so that meaningful temporal and interaction patterns can be captured, and, on the other, that it is constrained enough for this learning to be computationally feasible and mathematically well conditioned. The key premise of the proposed model—supported by evidence from previous research and validated by a series of experiments reported in this article—is that the future development of a patient's medical state can be predicted well by the presence of the diagnosis of a specific condition at some point in the patient's past. Thus I introduced a novel representation of a patient's health state in the form of a history vector—a fixed-length vector with binary values which correspond to the presence or absence of specific diagnoses in the patient's medical history. The model then learns the probabilities associated with semantically valid state transitions. The power of this model was demonstrated using a large, real-world data corpus collected by a local hospital, on which it was shown to outperform previous approaches in the literature, achieving over 91% accuracy in the prediction of the first future diagnosis for a patient given the patient's present medical history (in comparison with 35% accuracy achieved by the previous state-of-the-art). The proposed method was vastly superior for long-term prognosis as well, outperforming previous work in 82% of the cases, while producing comparable performance in the remaining 18% of the cases.

The framework described in this work opens a range of possibilities for future research and additional improvement. First, given the outstanding results obtained by modeling on the second granularity level of the ICD classification hierarchy, the performance of the proposed method should be examined using finer granularity codes (i.e. deeper ICD hierarchy levels) which would allow more specific predictions to be made. To facilitate such learning it is necessary to collect a larger data corpus. Second, the proposed model which as described here captures sequential diagnostic information can be readily extended to include temporal information as well. For example, this can be achieved by learning the parameters of suitable parametric descriptions of the probability density functions associated with transitions between pairs of history vectors.

Conflict of Interest: none declared.

References

Allaudeen,N. *et al.* (2011) Redefining readmission risk factors for general medicine patients. *J. Hosp. Med.*, **6**, 54–60.

Arandjelović,O. (2011) Contextually learnt detection of unusual motion-based behaviour in crowded public spaces. *Proc. Int. Symp. Comp. Inf. Sci.*, 403–410.

Arandjelović,O. (2015) Prediction of health outcomes using big (health) data. *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*

Axon,R.N. and Williams,M.V. (2011) Hospital readmission as an accountability measure. *JAMA*, **305**, 504–505.

Bartolomeo,N. *et al.* (2008) A Markov model to evaluate hospital readmission. *BMC Med. Res. Methodol.*, **8**, 23.

Bottle,A. *et al.* (2006) Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *J. R. Soc. Med.*, **99**, 406–414.

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

De Gaetano,A. *et al.* (2008) Mathematical models of diabetes progression. *Am. J. Physiol. Endocrinol. Metab.*, **295**, E1462–E1479.

Dharmarajan,K. *et al.* (2013) Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *JAMA*, **309**, 355–363.

Duffy,N.D. and Yau,J.F.S. (1995) 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**, 1531–1543.

Folino,F. and Pizzuti,C. (2011) Combining Markov models and association analysis for disease prediction. *Proc. Conf. Inform. Tech. Bio. Med. Inform.*, 39–52.

Friedman,B. *et al.* (2008–2009) Costly hospital readmissions and complex chronic illness. *Inquiry*, **45**, 408–421.

Gabriel,K.R. and Neumann,J. (1962) A Markov chain model for daily rainfall occurrence at Tel Aviv. *Q. J. R. Meteorol. Soc.*, **88**, 90–95.

Hammill,B.G. *et al.* (2011) Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circ. Cardiovasc. Qual. Outcome*, **4**, 60–67.

Holloway,J.J. *et al.* (1990) Risk factors for early readmission among veterans. *Health Serv. Res.*, **25**, 213–237.

Holman,C.D.A.J. *et al.* (2005) A multipurpose comorbidity scoring system performed better than the Charlson index. *J. Clin. Epidemiol.*, **58**, 1006–1014.

Jackson,C.H. *et al.* (2003) Multistate Markov models for disease progression with classification error. *J. R. Stat. Soc. Ser. D*, **52**, 193–209.

Kansagara,D. *et al.* (2011) Risk prediction models for hospital readmission: a systematic review. *JAMA*, **306**, 1688–1698.

Kilkenny,M.F. *et al.* (2013) Factors associated with 28-day hospital readmission after stroke in Australia. *Stroke*, **44**, 2260–2268.

Leegon,J. *et al.* (2005) Predicting hospital admission for emergency department patients using a Bayesian network. *AMIA Annu. Symp. Proc.*, 1022.

Li,J. and Guo,L. (2009) Hospital admission prediction using pre-hospital variables. *BIBM Conf. Proc.*, 283–286.

Mudge,A.M. *et al.* (2011) Recurrent readmissions in medical patients: a prospective study. *J. Hosp. Med.*, **6**, 61–67.

Sukkar,R. *et al.* (2012) Disease progression modeling using hidden Markov models. *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2845–2848.

Topp,B. *et al.* (2000) A model of β-cell mass, insulin, and glucose kinetics: Pathways to diabetes. *J. Theor. Biol.*, **206**, 605–619.

Vigod,S.N. *et al.* (2013) Within-hospital readmission: an indicator of readmission after discharge from psychiatric hospitalization. *Can. J. Psychiatry*, **58**, 476–481.

Wang,X. *et al.* (2014) Unsupervised learning of disease progression models. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data. Min.*, 85–94.

Whittaker,J.A. and Thomason,M.G. (1994) A Markov chain model for statistical software testing. *IEEE Trans. Softw. Eng.*, **20**, 812–824.

World Health Organization (2004) *International Statistical Classification of Diseases and Related Health Problems*. Vol. 1. World Health Organization, Geneva.

Ye,W. *et al.* (2012) Use of secondary data to estimate instantaneous model parameters of diabetic heart disease: lemonade method. *Inform. Fusion*, **13**, 137–145.