Towards Sophisticated Learning from EHRs: Increasing Prediction Specificity and Accuracy Using Clinically Meaningful Risk Criteria

Ieva Vasiljeva and Ognjen Arandjelović
School of Computer Science
University of St Andrews
St Andrews KY16 9SX
Fife, Scotland
United Kingdom

***Abstract*— Computer based analysis of Electronic Health Records (EHRs) has the potential to provide major novel insights of benefit both to specific individuals in the context of personalized medicine, as well as on the level of population-wide health care and policy. The present paper introduces a novel algorithm that uses machine learning for the discovery of longitudinal patterns in the diagnoses of diseases. Two key technical novelties are introduced: one in the form of a novel learning paradigm which enables greater learning specificity, and another in the form of a risk driven identification of confounding diagnoses. We present a series of experiments which demonstrate the effectiveness of the proposed techniques, and which reveal novel insights regarding the most promising future research directions.**

I. 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 [2], [4], [1]. 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 [6], [22], [11]. 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 work could deliver in terms of personalized medicine or health care policy [13] on the one hand, and a wiser utilization of resources [18], aid, and educational material [8], especially in the economically deprived countries [17], to appreciate the global and overarching potential.

II. Previous work

The contributions of the present work, the problems it addresses, and limitations of previous work that it overcomes are best understood in the context of a successful, recently described algorithm for longitudinal diagnosis pattern extraction from EHRs described by Arandjelović [5], [7] and subsequently further developed by Vasiljeva and Arandjelović [20]. Hence we summarize its main features; the reader is referred to the original publication for an in-depth description of the algorithm.

Consider a patient's hospital diagnosis history H which comprises a sequence of diagnoses d_i:

$$H = d_1 \rightarrow d_2 \rightarrow \ldots \rightarrow d_n, \tag{1}$$

where each d_i is a discrete variable whose value is a specific diagnostic code. A typical and widely used diagnosis coding scheme is that provided by the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [21]. The algorithm proposed in [6] predicts the most likely next diagnosis d^*_{n+1} for a patient by learning the probabilities of transitions from H to all other possible histories which can result from a single follow-up diagnosis d:

$$d^*_{n+1} = \arg\max_{d \in D} p(H \rightarrow d|H), \tag{2}$$

where D is the set diagnostic codes. To make the estimation of the probability $p(H \rightarrow d|H)$ tractable and learnable from limited data, a patient's diagnostic history H is represented using a fixed length binary vector $v(H)$. This representation bears resemblance to the bag of words representation frequently used in text analysis [9] and which has since been successfully adapted to various other application domains such as computer vision [3]. Each element in $v(H)$ encodes the presence (value 1) or lack thereof (value 0) of a specific salient diagnosis (i.e. the corresponding diagnostic code) in H, save for the last element which captures jointly all non-salient diagnoses. Saliency is determined by the frequency of the corresponding diagnosis in the entire data corpus (n.b. different saliency criteria can be readily used instead). The probability $p(H \rightarrow d|H)$ in (2) is then estimated by superimposing a Markovian model [19], [15] on the space of history vectors which leads to $H \rightarrow d$ being interpreted as a transition from the state represented by $v(H)$ to the state represented by $v(H \rightarrow d)$. As usual the probabilities parameterizing the Markov model are learnt from a training data corpus. A conceptual illustration of the method is shown in Fig 1.

The key idea behind the described model is that it is the *presence* of past complications which most strongly predicts future ailments [16], [14], [12], [10], which allows

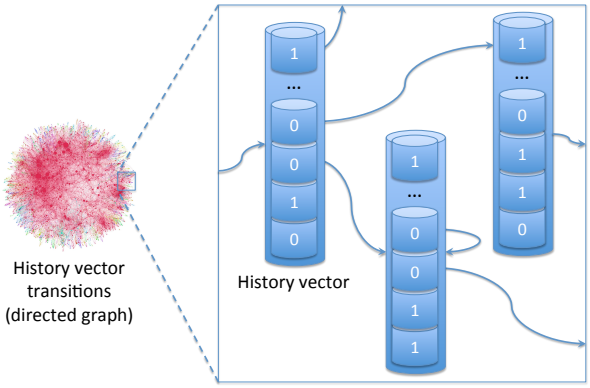

Fig. 1. Conceptual illustration of the method proposed in [6] which superimposes a Markovian model over a space of history vectors used to represent the medical state of a patient.

for the space of states over which learning is performed to be reduced dramatically; in particular, this is achieved by employing a fixed length state representation and through binarization of its elements.

III. 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*. 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 the 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.

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 expect 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:

$$\rho = \rho_x + \rho_y. \tag{3}$$

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:

$$\rho_x = C_x \int_z^1 |x - z| p(x|n_x) dx + \tag{4}$$

$$+ (1 - C_x) \int_0^z |z - x| p(x|n_x) dx. \tag{5}$$

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

$$p(x|n_x) = \frac{p(n_x|x)p(x)}{p(n_x)}, \tag{6}$$

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:

$$p(x|n_x) = \frac{p(n_x|x)}{p(n_x)} = \frac{\binom{N_x}{n_x} x^{n_x}(1-x)^{N_x-n_x}}{\int_0^1 p(n_x|w)dw} \quad (7)$$

$$= \frac{\binom{N_x}{n_x} x^{n_x}(1-x)^{N_x-n_x}}{\int_0^1 \binom{N_x}{n_x} w^{n_x}(1-w)^{N_x-n_x} dw} \quad (8)$$

$$= \frac{x^{n_x}(1-x)^{N_x-n_x}}{\beta(n_x+1, N_x-n_x+1)} \quad (9)$$

where $\beta(.)$ is the Euler beta function, and simple marginalization over x is performed in the denominator. This expression can be substituted back into (4) and (5), and then (3), and the integration performed numerically (which is both simple and fast, given that it is a simple integration in 1D). Merging is then performed if the weighted proportion of incorrect predictions exceeds a certain threshold t_m, set e.g. by a physician.

a) 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 can is also readily interpretable. For example, for a terminal diagnosis the integrand in (4) 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, for start. Similarly, for a terminal diagnosis the integrand in (5) estimates the number of individuals who would experience a terminal episode but 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 clinical 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.

IV. 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.

A. 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 conducted our experiments the large collection of EHRs (over 40,000 individuals and over 400,000 diagnostic events) described in [6]. For completeness here we summarize the key features of this data set.

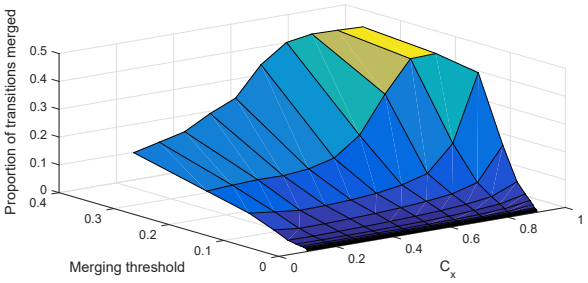

Fig. 2. Surface plot showing the number of pair-wise merges performed (as the proportion of all possible transitions pairs which could possibly be merged) as a function of the adjustable parameters of the proposed method, namely the merging threshold t_m and the relative risk weighting constant C_x in (4) and (5).

The EHRs adopted for evaluation were collected by a large private hospital. The distribution of patient age in the database is 73 ± 15 years, the youngest and oldest patients being 17 months and 102 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 10.1 ± 62.2.

B. Experiments, results, and discussion

Using the real-world collection of EHRs described in the previous section, we conducted a series of experiments to facilitate the understanding of the proposed merging technique.

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 (4) and (5). 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 Fig 2. 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 Sec III) 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

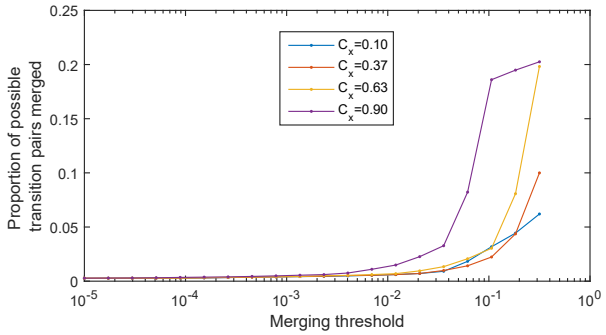

Fig. 3. The number of effected merges associated with the diagnosis of stroke (as d_f in Sec III) as the proportion of all possible transitions pairs which could possibly be merged and associated with transitions effected by the diagnosis of stroke.

behaviour in more detail. Interpreted intuitively, this means that on average the diagnosis of stroke affects the least (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 Fig 3. It is insightful to observe that much like in Fig 2, 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.

V. Summary and Future Work

In this paper we introduced a novel, clinically informed method for improving a previously described algorithm that uses machine learning on EHR collections for the discovery of longitudinal patterns in the diagnosis of diseases. The key technical novelty comes in the form of risk driven identification of confounding diagnoses which allows for better utilization of available data and more reliable prediction. Experiments on a large real-world data corpus of EHRs were used to analyse the performance of the proposed technique.

As regards possible future work directions, a number of possibilities were highlighted in the work which originally introduced the history vector based approach [5]. Our work, both previous [20] and that described in the present paper, provides additional evidence that the aforementioned possibilities are promising, while suggesting a number of potentially more significant immediate alternatives. In particular while we agree with the suggestion in the original paper that the presence of a particular diagnosis is a predictive factor not much weaker than the exact count of the same diagnosis (the use of which would likely require prohibitively large amounts of training data), 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 diagnoses or 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.

References

[1] 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*, 2016.

[2] O. Arandjelović. A new framework for interpreting the outcomes of imperfectly blinded controlled clinical trials. *PLOS ONE*, 7(12):e48984, 2012.

[3] O. Arandjelović. Object matching using boundary descriptors. *In Proc. British Machine Vision Conference*, 2012. DOI: 10.5244/C.26.85.

[4] O. Arandjelović. Clinical trial adaptation by matching evidence in complementary patient sub-groups of auxiliary blinding questionnaire responses. *PLOS ONE*, 10(7):e0131524, 2015.

[5] O. Arandjelović. Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, 31(24):3970–3976, 2015.

[6] 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, August 2015.

[7] O. Arandjelović. On the discovery of hospital admission patterns – a clarification. *Bioinformatics*, 2016. DOI: 10.1093/bioinformatics/btw049.

[8] 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.

[9] 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, 2015.

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

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] D. Scanfeld, V. Scanfeld, and E. L. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188, 2010.

[19] 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.

[20] 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*, 2016.

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

[22] 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.