

Intuitive and Interpretable Visual Communication of a Complex Statistical Model of Disease Progression and Risk

Jieyi Li and Ognjen Arandjelović
School of Computer Science
University of St Andrews
Scotland, United Kingdom

Abstract—Computer science and machine learning in particular are increasingly lauded for their potential to aid medical practice. However, the highly technical nature of the state of the art techniques can be a major obstacle in their usability by health care professionals and thus, their adoption and actual practical benefit. In this paper we describe a software tool which focuses on the visualization of predictions made by a recently developed method which leverages data in the form of large scale electronic records for making diagnostic predictions. Guided by risk predictions, our tool allows the user to explore interactively different diagnostic trajectories, or display cumulative long term prognostics, in an intuitive and easily interpretable manner.

I. INTRODUCTION

With an increased availability of digitalized medical information (scans, records of diagnoses, prescriptions, and treatments, etc.) recent years have been witnessing an increasing amount of research attention on the use of modern computer science – namely, machine learning, data mining, pattern recognition, and others – in medicine and health care. These efforts range greatly in nature and include those which focus on improving diagnosis [13], incentivizing behavioural change [12], [15], measuring the penetration of health care advice [6], assessing provider-patient interaction quality [4], and numerous others [8]. Considering that there is a major technical (and, arguably, methodological) gap between health care professionals and computer scientists, the question of how machine extracted information is communicated is crucial [5]. In some applications this is easier than in others. For example, a decision on the diagnosis is usually quite simple to convey (though the process through which a machine reached this decision may not be, which can be an important issue in its own right as well) [10]. However, certain types of information are much more difficult to communicate in a manner which is readily understood by individuals who are not experts in computer science. Indeed, in the present paper our aim is to bridge this representational gap in the context of a recently developed algorithm which uses large amounts of electronic medical records to predict the probability of future diagnostic events.

II. UNDERLYING MODEL

For completeness herein we present a summary of the key ideas of the adopted method. For in-depth technical details, and the related discussion and results, the reader is referred to the original publication [1] as well as the body of related follow-up work [18], [16], [19].

A. Electronic medical records

Large amounts of medical data are 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 diagnosed with an ailment. In particular, the diagnosis is recorded using a standardized coding schema such as that provided by the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [20] or the related Australian Refined Diagnosis-Related Groups (AR-DRGs), which have hierarchical structures [3].

B. Underlying model

The predictions the present work aims to visualize are underlain by the epidemiological predictions of the model originally introduced by Arandjelović [2]. The model predicts the probability of a specific diagnosis a following the patient history H :

$$p(H \rightarrow a|H). \quad (1)$$

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. Consider a particular diagnosis history $H = d_1 \rightarrow \dots \rightarrow d_n$. The adopted method makes use of the observation that when it comes to chronic diseases, the very *presence* of past complications strongly predicts future complications [11], [9]. Thus, a history H is represented using a history vector $v = v(H)$ which is a fixed length vector with binary values. Though seemingly crude this representation has been used with success in other domains too [7]. 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:

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

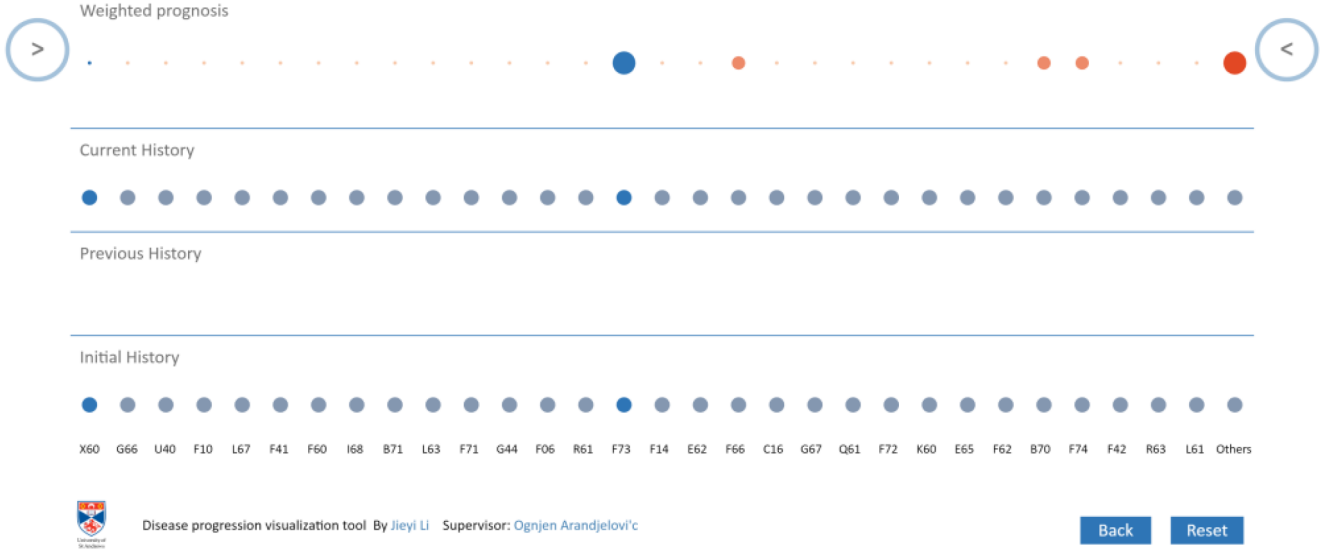


Fig. 1. The main window of our visualization tool.

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 predicated 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:

$$p(v(H) \rightarrow v(H')). \quad (2)$$

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:

$$v(H)_{i(d)} = 1 \wedge v(H')_{i(d)} = 0 \Rightarrow p(v(H) \rightarrow v(H')) = 0. \quad (3)$$

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. 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) 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 is evidenced by statistics which show that only a small number of diagnosis types covers a vast amount of all data. For example the top 30 most frequent types account for 75% of all diagnoses.

III. VISUALIZATION

There are several noteworthy aspects of the data and the model summarized in the preceding section which make visualization difficult. Firstly, even if a reduced set of diagnoses is considered, the number of diagnoses of interest and their potential relationships already pose a challenge. An approach which highlights salient information and which allows the user control over how information is organized is highly desirable. Secondly, much of the inner workings of the model, which are mathematical, are not readily understandable by a non-specialist. Yet they have a profound significance. Being able to display probabilistic information, especially when it exhibits itself in a cumulative manner over simulated time, is crucial in making the system useful to health care providers.

In the next section we start by laying out the main design of our application and explaining the key visualization tools we employ. The main functionality of the application

captured by this section allows the health care practitioner to visualize immediate risk and explore different disease progression alternatives in a simple manner. This section is followed up by sections which detail a series of useful extensions to the basic functionality. These allow for the visualization of long term risk using different prediction paradigms.

A. Main visualization framework

As explained in the previous section, the cornerstone representation in the model which we build our visualization around comprises vectors with binary entries. This conceptually simple representation allowed us to come up with an elegant design which immediately draws the user’s attention to salient features in a patient’s medical history. The main window of our application is shown in Figure 1. To start, consider the bottommost row of filled circles. Each circle corresponds to a diagnosis included in the predictive model, as indicated by the corresponding diagnostic code underneath. Notice that the only aspect in which the appearance of a circle can vary is its colour. In particular, we denote diagnoses present in a patient’s history using dark blue and those which are not present using light blue.

Next, observe that there are multiple histories displayed concurrently. The bottommost history, labelled ‘Initial History’, corresponds to the history from which the space of possible diagnostic trajectories is explored. In clinical practice this initial history will usually be the diagnostic record of a patient at admission. Thereafter exploration proceeds by the user selecting a specific diagnosis (by clicking the corresponding circle). This action changes the history denoted ‘Current History’ which corresponds to the current state in the exploratory process and is guided by information in the topmost row. Unlike the three other rows which display the same type of information, namely diagnostic histories, the circles in this row also vary in their size and colour which encode the probability of a specific diagnosis given the current diagnostic history, estimated using the model detailed in the previous section. Thus, the user is informed in the exploratory process and can pursue possible diagnostic futures which are more likely.

B. Selective emphasis

Recall that the original work which introduced the adopted prediction model based the diagnostic history vector on the 30 most common diagnoses. This is a sufficiently low number to allow for the visualization described in Section III-A to appear uncluttered on most devices. However, subsequent work has demonstrated that the model is successful even with the inclusion of a much greater number of diagnoses which can be of clinical interest [17]. Attempting to visualize these in the same manner clearly poses problems with clutter and the ease with which diagnoses of interest can be observed. To overcome this obstacle we came up with two solutions. Firstly, we allow the user to select or deselect specific diagnoses from being visualized. Deselected diagnoses are still included in the predictive model but their states are not displayed in the main window. Secondly, we make use of the hierarchical nature of diagnostic coding. In

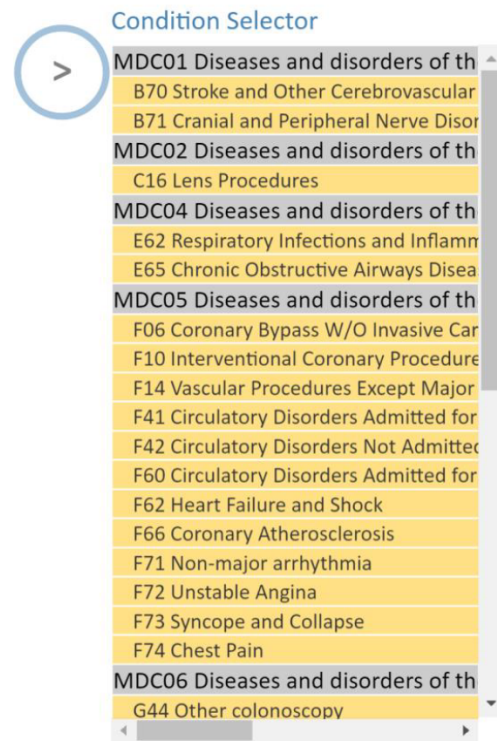


Fig. 2. Selector of diagnoses of interest. Diagnoses can be selected or deselected for display purposes, or grouped according to the hierarchy of the used diagnostic coding schema.

particular, our application supports several common coding schemes, including ICD-10 and AR-DRG, and thus allows for diagnoses to be grouped according to the subtree in the hierarchy. In other words, rather than displaying related diagnoses separately, the user can choose to unify these and visualize merely that any of the diagnoses of a specific group are present. As before, if more granularity is desired at any point, the option can be changed and individual diagnoses displayed, given that it is only the visualization which is altered and not the underlying predictive model.

C. Additional prognostics

The exploration of diagnostic futures described thus far is local in the sense that the user can see the predictions of short term risks and using this information make incremental moves through the tree of different possibilities. However, considering that the original work has demonstrated good performance on the task of long term prediction, we also sought ways of visualizing this aspect of the model too. This information is useful in that it can be time saving, more incentivizing to patients, and direction providing in incremental exploration. Hence, we provide the option to display an additional type of prediction. In particular, we sample ultimate diagnostic histories reached from the current history and display the probability of each diagnosis according to the proportion of ultimate histories in which it appears. As before, probabilities are encoded using size and colour, as shown in Figure 3.

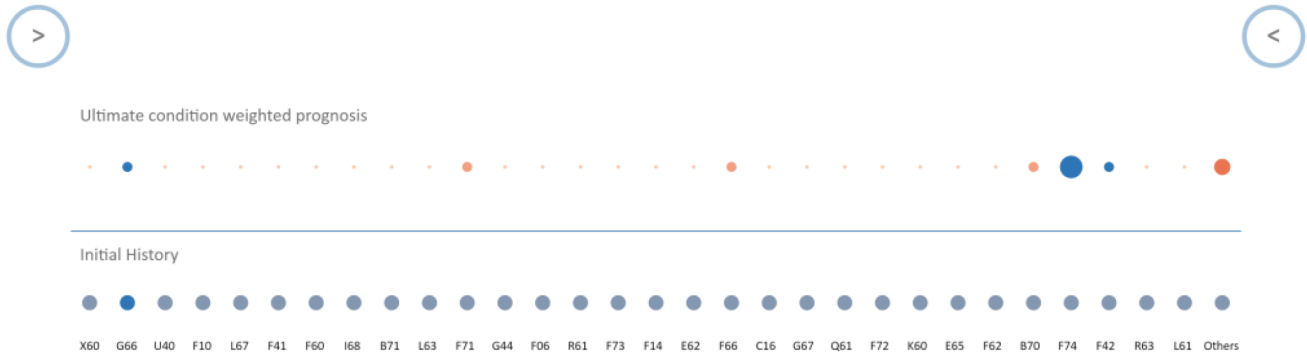


Fig. 3. Ultimate condition weighted prognosis shows the proportion of ultimate histories in which different diagnoses appears.

IV. SUMMARY AND FUTURE WORK

In this paper we described a visualization tool capable of displaying the predictions of a recently introduced machine learning based algorithm shown to be effective at predicting the risk of different ailments based on a patient's current diagnostic history. Our visualization approach makes use of several key properties of the adopted algorithm and the structure of data. The binary history vector at the heart of the adopted model naturally led us to an intuitive colour-coding of the presence of a specific diagnosis and the exploration of different future progressions by simple mouse clicks. On the other hand, the hierarchical form of common diagnosis coding schemas allows us to keep the visualization compact and clutter-free, by including options to group diagnoses by hierarchical similarity and by facilitating the hiding of diagnoses which are not of interest. In addition to interactive incremental exploration of a patient's state, our tool also includes the possibility of showing long term prognostics based on probabilistic sampling of different progression trajectories. Our current research efforts which build upon the present contribution are aimed at increasing the amount of visualized information – and the inclusion of temporal (rather than merely sequential) statistics – while retaining simplicity of use and ease of comprehension by health care professionals [14].

REFERENCES

- [1] O. Arandjelović. Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, 31(24):3970–3976, 2015.
- [2] 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, 2015.
- [3] O. Arandjelović. On the discovery of hospital admission patterns – a clarification. *Bioinformatics*, 32(13):2078, 2016.
- [4] L. Barracliff, O. Arandjelović, and G. Humphris. Can machine learning predict healthcare professionals' responses to patient emotions? *In Proc. International Conference on Bioinformatics and Computational Biology*, pages 101–106, 2017.
- [5] J. R. Bautista and T. T. Lin. Sociotechnical analysis of nurses' use of personal mobile phones at work. *International Journal of Medical Informatics*, 95:71–80, 2016.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] J. Butler and A. Kalogeropoulos. Hospital strategies to reduce heart failure readmissions. *J Am Coll Cardiol*, 60(7):615–617, 2012.
- [10] P. D. Caie and D. J. Harrison. Next-generation pathology. *Systems Medicine*, pages 61–72, 2016.
- [11] 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.
- [12] J. Li and O. Arandjelović. Glycaemic index prediction: a pilot study of data linkage challenges and the application of machine learning. *In Proc. IEEE International Conference on Biomedical and Health Informatics*, pages 357–360, 2017.
- [13] A. Oniško and M. J. Druzdzel. Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artif Intell Med*, 57(3):197–206, 2013.
- [14] R. Osuala and O. Arandjelović. Visualization of patient specific disease risk. *In Proc. IEEE International Conference on Biomedical and Health Informatics*, pages 241–244, 2017.
- [15] R. S. Sadasivam, S. L. Cutrona, R. L. Kinney, B. M. Marlin, K. M. Mazor, S. C. Lemon, and T. K. Houston. Collective-intelligence recommender systems: advancing computer tailoring for health behavior change into the 21st century. *Journal of Medical Internet Research*, 18(3):e42, 2016.
- [16] 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*, 2016.
- [17] 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, 2016.
- [18] 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, 2016.
- [19] I. Vasiljeva and O. Arandjelović. Diagnosis prediction from electronic health records (EHR) using the binary diagnosis history vector representation. *Journal of Computational Biology*, 2017.
- [20] World Health Organization. *International statistical classification of diseases and related health problems.*, volume 1. World Health Organization, 2004.