Visualization of Patient Specific Disease Risk Prediction

Richard Osuala and Ognjen Arandjelović
University of St Andrews, United Kingdom

***Abstract*— The increasing trend of systematic collection of medical data (diagnoses, hospital admission emergencies, blood test results, scans etc) by health care providers offers an unprecedented opportunity for the application of modern data mining, pattern recognition, and machine learning algorithms. The ultimate aim is invariably that of improving outcomes, be it directly or indirectly. Notwithstanding the successes of recent research efforts in this realm, a major obstacle of making the developed models usable by medical professionals (rather than computer scientists or statisticians) remains largely unaddressed. Yet, a mounting amount of evidence shows that the ability to understanding and easily use novel technologies is a major factor governing how widely adopted by the target users (doctors, nurses, and patients, amongst others) they are likely to be. In this work we address this technical gap. In particular, we describe a portable, web based interface that allows health care professionals to interact with recently developed machine learning and data driven prognostic algorithms. Our application interfaces a statistical disease progression model and displays its predictions in an intuitive and readily understandable manner. Different types of geometric primitives and their visual properties (such as size or colour), are used to represent abstract quantities such as probability density functions, the rate of change of relative probabilities, and a series of other relevant statistics which the heath care professional can use to explore patients' risk factors or provide personalized, evidence and data driven incentivization to the patient.**

I. Introduction

Electronic medical records (EMRs), nowadays a routinely collected data resource in hospitals in economically developed countries, offer an exciting opportunity for machine learning based knowledge discovery which could significantly affect health care delivery, its quality, and therefore intervention outcomes [18], [8], [11]. Some of the most prominent problems addressed by the existing literature include the discovery of risk factors, the modelling of disease progression patterns, and the development of patient specific prognostics [4], [6], [16]. However, a major challenge posed by the need to interface these technological advancements with medical personnel and patients themselves, has attracted much less research attention [9], [12], [3]. Yet, some of the very premises of the work on person specific prognosis include the incentivization of patients [2]. Moreover, the ability to interact with technology in an intuitive manner is a major aspect governing its adoptability in actual health care practice [5], [13].

The visualization tool we introduce in this work is built around a recently proposed disease progression model which has demonstrated highly promising results on real world data [2], [16]. This model, and indeed all models likely to be successful on the task of comorbidity modelling and prediction, is highly technical and in that sense not readily accessible to medical practitioners or patients. A large volume of previous work has shown that this can be a major obstacle in the adoption of technology in the clinical context [5], [9]. Thus, the contribution of this paper is a novel framework which makes a major step towards bridging this gap of outstanding practical significance.

II. Under the hood: the underlying prediction 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 publications [1], [15], [16], [14].

The history vector based sequential prediction model we adopt from the work of Arandjelović [2] treats a patient's medical record as comprising a sequence of hospital admissions $a_1, \ldots, a_i, \ldots, a_n$ which form a hospital admission history H:

$$H = a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow a_n \tag{1}$$

Each a_i is a discrete event coded using one of a number of standard disease coding schemas e.g. [17] or one of a number of mostly related alternatives [7]. The most likely follow-up admission a^*_{n+1} is calculated by likelihood maximization from the current history:

$$a^*_{n+1} = \arg\max_a p(H \rightarrow a) \tag{2}$$

The method proposed by Arandjelović represents a history as a fixed length binary history vector $v = v(H)$ over the most common disease diagnoses, where 1 denotes the presence of a specific diagnosis in the history, and 0 absence thereof. The transition probabilities between different history vectors $p(v(H_1) \rightarrow v(H_2))$ are learnt from a training data corpus.

The original model described in [2], [16] facilitates sequential prediction only. In other words, it predicts the next diagnosis for a patient (or, equivalently, provides a probability ranked list of diagnoses) without any associated temporal information i.e. it is not able to predict the timing of this diagnosis. Herein the original model is further endowed with a temporal predictive ability. This is achieved by learning the cumulative distribution function (cdf) of a transition from one history vector to another. Considering that a appropriate probability density function (pdf) associated with transitions is the log-normal distribution:

$$p_t(t) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\ln(t) - \tau}{\sigma\sqrt{2}}\right)\right] \tag{3}$$

the corresponding cdf is:

$$P_t(x) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{-(\ln t - \tau)^2}{2\sigma^2}} \tag{4}$$

where t is the temporal distance of the transition measured from the present, and τ and σ the parameters of the distribution, frequently referred to as the 'location' and 'scale' parameters. The two parameters are also readily learnt from the training data corpus using standard maximum likelihood estimation.

III. Visualization of Model Predictions

By consulting with a number of relevant health care professionals (clinicians, doctors, and nurses) and by adopting an iterative design-test-reassess design process, we found that different users found different manners of information presentation most intuitive and easiest to understand. Consequently we developed a combination of different visualization options which can be readily switched between by the user.

A. Blob based visualization

The first circle based visualization approach resembles a so-called blob chart [10], with equidistant blobs which represent different disease diagnoses being distributed horizontally, as illustrated in Fig 1. The corresponded diagnoses are labelled using their codes under the adopted coding system (e.g. WHO's diagnosis related groups [17], or the Australian refined diagnosis-related groups [7]). These are standard codes, used widely and understood by health care professionals, and allow for the diagnoses to be shown in a succinct, clutter free manner. Additional information and a more detailed description of a diagnosis can be obtained by clicking any visualization element associated with the diagnosis (its label or the corresponding blob, in this visualization).

The size of a particular blob encodes the value of the cumulative density function corresponding to the occurrence of the respective diagnosis by the specific time in future. This time is specified by the user and allows the user to gain an understanding of the highest risks for the patient within this period. Larger and thus more prominent blobs (and hence the corresponding diagnoses) draw the user's attention to the most probable complications while at the same time providing a simple way of judging relative risks too – several large blobs immediately suggest a cluster of comorbidities, whereas single dominant blob highlights a specific primary diagnosis of interest.

Moreover, we encode the rate of the cdf change by a blob's colour, using the standard heat map. In this manner, in addition to the instantaneous value of the cdf, we communicate to the user the possibly uneven changes in the probabilities of different diagnoses over time. By including this information in our visualization, a clinician can be alerted of a high risk increase in the near future (relative to the currently selected date of interest). The blob chart visualization is set as the default visualization, whereas the other two visualization options (described next), if selected, are displayed in modal windows.

B. Bar chart based visualization

The second visualization visualization option uses the well known bar chart encoding with the height of bars representing the corresponding value of the cumulative distribution function at the specific date of interest, as illustrated in Fig 2. As before, bar colours communicate the rate of change of the cumulative distribution function across time and the smaller rectangles underneath the bars represent the presence (or lack thereof) of different diagnoses in the current history vector. Fundamentally this visualization conveys the same information as the other two alternatives, namely the blob based and radial chart based visualizations shown respectively in Figs 1 and 3, but its different way of encoding this information was found to be preferred by some users. Hence we found it to be a useful alternative to include.

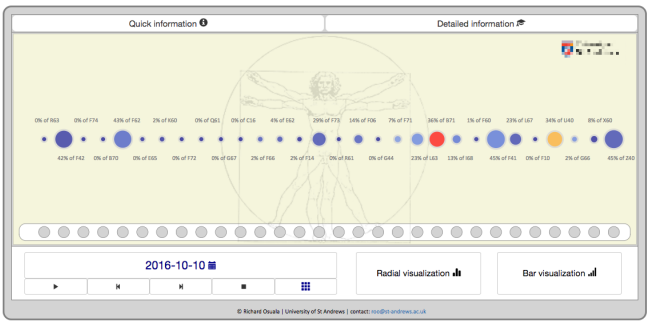

Fig. 1. Blob chart visualization is the default visualization in our application. Each blob represents a disease diagnosis with the blob size encoding the value of the corresponding probability density function at the selected instance in time. In the example shown in this figure all history vector entries are set to 0, indicating the absence of any diagnoses in the patient's medical history.

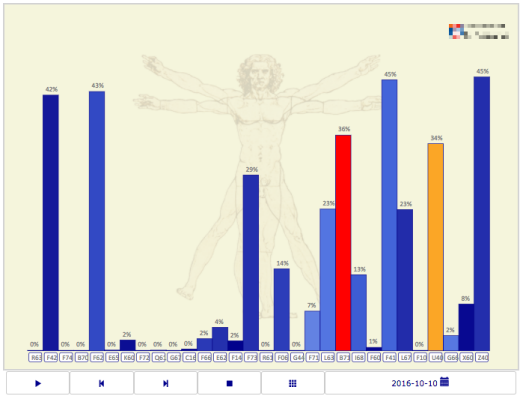

Fig. 2. Bar chart visualization was found to be preferred by some users. Fundamentally it conveys the same information as the other two alternatives, shown in Figs 1 and 3 but differently encoded.

C. Radial chart based visualization

The third and final visualization we developed resembles a radial chart with rectangles spreading out radially, as shown in Fig 3. During the course of our interviews with the target users we found that some of them preferred this layout to the two described previously due to its symmetry – the lack of symmetry in the first two visualizations suggested to some users some differentiation between different diagnoses which is neither intended nor present in the underlying method or its output. As with the previous visualization, the radial length and colour of rectangles are used to represent the value of

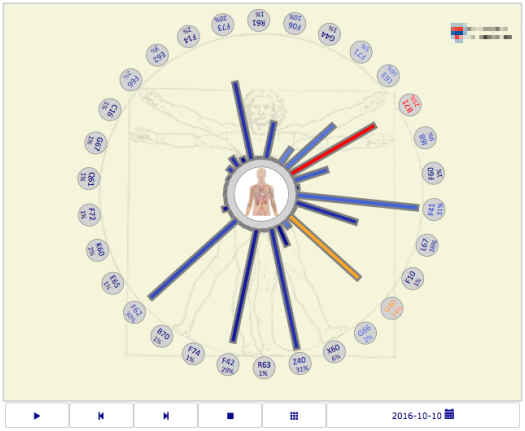

Fig. 3. Radial chart visualization was found to be preferred by some users. Fundamentally it conveys the same information as the other two alternatives, shown in Figs 1 and 2 but differently encoded.

the corresponding cumulative distribution functions and their rates of change. We found that with this visualization the users also clearly associated the circles that represent the binary history vector entries (i.e. the presence of specific diagnoses in a patient's EMR history) with the corresponding diagnoses.

D. Interactive features

In all of the visualizations, by select on any cdf encoding element the user can open a window which displays further detailed information and allows for the status of the diagnosis to be changed, as illustrated in Fig 4. A fast way of flipping the status of a diagnosis (present or not present) is also provided – a user can simply click on the green (add) or the red (remove) buttons. This effects a history vector transition which in turn triggers a change in the corresponding visual representations (e.g. blob size and colour). This feature allows the health care practitioner to explore how different potential diagnoses (e.g. those that the patient may be at the greatest risk of developing) affect the patient's health state further in the future. This can be used as a powerful incentivization tool. For example, the patient can be shown how a specific ailment that he/she is at the risk of developing due to lifestyle choices (e.g. smoking, excessive food intake etc), would influence other health related outcomes (e.g. lung cancer, diabetes, hypertension etc).

E. Help, hints etc.

To facilitate instantaneous help and a ready understanding of different visual elements in our visualizations, when the cursor hovers over any of the relevant geometric entities, all associated information is emphasised. For example, as illustrated in Fig 4, after hovering over the blob representing the first disease included explicitly in the model [2], a line connecting the blob with the corresponding value of the probability density function is shown. Furthermore, in addition to the disease code, a full description of the diagnosis is shown both above the cursor and at the bottom of the window. To avoid so-called 'change blindness', further animations emphasise transitions between history vectors or changes

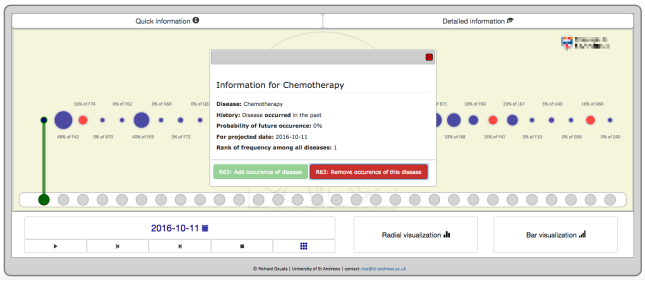

Fig. 4. In this example, a chemotherapy diagnosis is present in the patient's history vector. By selecting the corresponding blob, the user can open a modal window from which the diagnosis status can be changed. A line connecting the value of the probability density function corresponding to the diagnosis and the date selected (2016-10-11 in this case) is shown so as to emphasise this information.

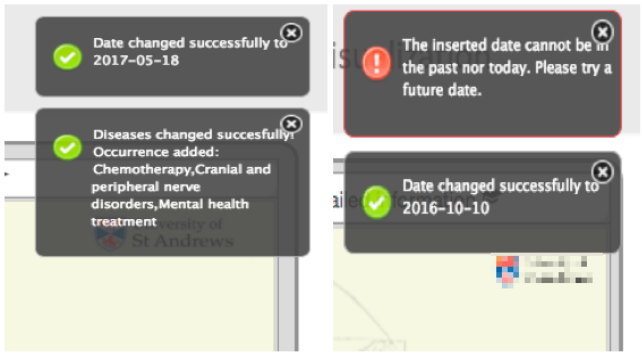

Fig. 5. Examples of notification messages displayed as feedback.

to the date of interest. This feedback uses highlighting and increased contrast against the beige coloured background of the current visualization after 0.5 seconds.

To minimize the chance of human error and assist the user in interaction, input checking and succinct, timely, and informative notification messages inform users of their interactions with our application. Notifications appear for example in case of validation errors (e.g. date needs to be in the future) or if the date or the history vector are changed due to an added diagnosis, as illustrated with a few examples in Fig 5. The selected date of interest is visible inside an interactive button below the visualization that can be clicked to make adjustments. After clicking on the button, a modal window is opened which allows the user to change the date using the familiar calendar view.

F. Automatic time lapse and long term outcome simulations

Our application also provides further interactive features, activated using buttons placed below the main visualization space. These buttons resemble the widely known and hence intuitively understandable functions of a media player, such as 'play', 'pause', 'stop' etc. These buttons provide effortless navigation through *time* via simulations of possible temporal trajectories through the space of possible diagnoses. Temporal transitions predicted by the adopted model are accompanied by the automatic visualization of the corresponding disease progression. The forward and backward buttons allow for manual time jumps. Such time jumps change the date of interest and update the visualization accordingly. The

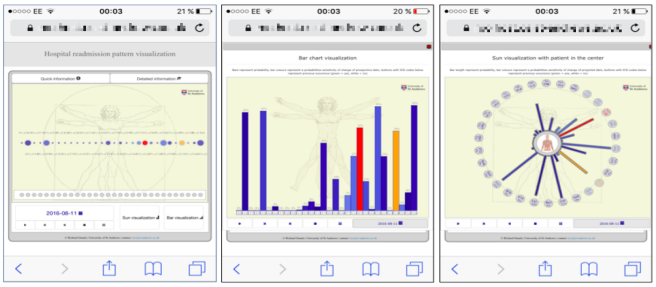

Fig. 6. Three visualizations as seen on a 3.5-inches screen in Mobile Safari 9 in an IOS 9.3.1 environment on a cell (mobile) phone.

duration of such time jumps (e.g. days, months or years) can be specified in the 'date selection' modal window.

Clicking the play button opens a modal window where users can also choose the length of time jumps, the real time between predicted transitions (e.g. every 2 seconds), and whether diagnoses should be added automatically upon exceeding a certain probability of occurrence (i.e. the corresponding cdf value). In the latter case, the play function can add future diagnoses deterministically by using maximum likelihood prediction or non-deterministically by pdf weighted random sampling (ensuring that more likely diagnostic paths are simulated with the correspondingly higher frequency). Once the play function is activated in the modal window, our application repeatedly makes forward temporal jumps (as explained earlier, their duration can be set by the user). If deterministic prediction is selected, diagnoses are added to the visualized medical history if the corresponding cdf exceeds a probability threshold which too can be specified by the user. Random sampling adds diagnoses using cdf based weighting, thus allowing the clinician to explore multiple future disease progression patterns with repeated activation of the function. When the play function is running, for clarity the 'play' button disappears, and is replaced by the 'pause' button. The click event of the pause button puts the play function on hold to enable users to explore the currently displayed simulated health care record in detail. Clicking the stop button terminates the play function, and resets the date of interest to its default value (the present date).

G. Note on implementation

Our visualization was implemented as a web application using the D3 Javascript library d3.js thereby offering high portability across different devices and operating system environments; see Fig 6 for an example. Additional advantages offered by its web based implementation include the simplicity of deployment, as no installation or configuration is needed, and an immediate sense of familiarity for non-technical users.

The d3.js based circles and rectangles used to visualize blobs and bars are nested in a scalable vector graphics (svg). Their radii and lengths are calculated using d3.js scale functions. Heat map colouring uses chroma.js interpolation between four plain colours and scaling with d3.js to calculate the corresponding mapping between the pdf rate of change values and the computed colour palette. To switch from the default blob chart to another visualization format, JQueryUI based modal functions append HTML code to the interface.

IV. Summary

In this paper we introduced an intuitive visual interface built around a recently proposed computational model of disease progression, aimed at making the model's predictions accessible to health professionals in their daily work. A range of interactive features allows the user to explore patient specific risk across time. To the best of the authors' knowledge, this is the first attempt at bridging the gap between increasingly complex machine learning based algorithms and the realm of heath care practice. We trust that our contribution will facilitate increased adoption of technology in health care delivery, empowering both the medical community and patients in understanding risk and how to address it. Moreover, we hope that our work will inspire future research in this realm.

References

[1] O. Arandjelović. Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, 2015.

[2] O. Arandjelović. Prediction of health outcomes using big (health) data. *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015.

[3] L. Barracliffe *et al.* Can machine learning predict healthcare professionals' responses to patient emotions? *International Conference on Bioinformatics and Computational Biology*, 2017.

[4] N. Bartolomeo *et al.* A Markov model to evaluate hospital readmission. *BMC Medical Research Methodology*, 2008.

[5] J. R. Bautista and T. T. Lin. Sociotechnical analysis of nurses' use of personal mobile phones at work. *International Journal of Medical Informatics*, 2016.

[6] 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. *Statistics in Medicine*, 1995.

[7] C. Kobel *et al.* DRG systems and similar patient classification systems in europe. *Diagnosis-Related Groups in Europe: moving towards transparency, efficiency and quality in hospitals, 1st edition, Open University Press and WHO Regional Office for Europe, Buckingham*, 2011.

[8] E. C. Lau *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. *Clinical Epidemiology*, 2011.

[9] T. Le *et al.* Health providers' perceptions of novel approaches to visualizing integrated health information. *Methods of Information in Medicine*, 2013.

[10] T. Munzner. *Visualization Analysis and Design.* CRC Press, 2014.

[11] P. M. Nadkarni. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. *Journal of the American Medical Informatics Association*, 2010.

[12] C. B. Nielsen. Visualization: a mind – machine interface for discovery. *Trends in Genetics*, 2016.

[13] C. Thuemmler *et al.* A methodology to assess social technological alignment in the health domain. *IRBM*, 2015.

[14] I. Vasiljeva and O. Arandjelović. Automatic knowledge extraction from EHRs. *IJCAI Workshop on Knowledge Discovery in Healthcare Data*, 2016.

[15] I. Vasiljeva and O. Arandjelović. Prediction of future hospital admissions – what is the tradeoff between specificity and accuracy? *International Conference on Bioinformatics and Computational Biology*, 2016.

[16] I. Vasiljeva and O. Arandjelović. Towards sophisticated learning from EHRs: increasing prediction specificity and accuracy using clinically meaningful risk criteria. *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016.

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

[18] S.-M. Zhou *et al.* Defining disease phenotypes in primary care electronic health records by a machine learning approach: a case study in identifying rheumatoid arthritis. *PLOS ONE*, 2016.