Predicting Within-24h Visualisation of Hospital Clinical Reports Using Bayesian Networks

Pedro Pereira Rodrigues\textsuperscript{1,2,3(✉)}, Cristiano Inácio Lemes\textsuperscript{4}, Cláudia Camila Dias\textsuperscript{1,2}, and Ricardo Cruz-Correia\textsuperscript{1,2}

\textsuperscript{1} CINTESIS - Centre for Health Technology and Services Research, Rua Dr. Plácido Costa, s/n, 4200-450 Porto, Portugal \{pprodrigues,camila,rcorreia\}@med.up.pt
\textsuperscript{2} CIDES-FMUP, Faculty of Medicine of the University of Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
\textsuperscript{3} LIAAD - INESC TEC, Artificial Intelligence and Decision Support Laboratory, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
\textsuperscript{4} ICMC, Institute of Mathematical and Computer Sciences, University of São Paulo, Avenida Trabalhador São-carlense, 400, São Carlos 13566-590, Brazil cristianoinaciolemes@gmail.com

Abstract. Clinical record integration and visualisation is one of the most important abilities of modern health information systems (HIS). Its use on clinical encounters plays a relevant role in the efficacy and efficiency of health care. One solution is to consider a virtual patient record (VPR), created by integrating all clinical records, which must collect documents from distributed departmental HIS. However, the amount of data currently being produced, stored and used in these settings is stressing information technology infrastructure: integrated VPR of central hospitals may gather millions of clinical documents, so accessing data becomes an issue. Our vision is that, making clinical reports to be stored either in primary (fast) or secondary (slower) storage devices according to their likelihood of visualisation can help manage the workload of these systems. The aim of this work was to develop a model that predicts the probability of visualisation, within 24h after production, of each clinical report in the VPR, so that reports less likely to be visualised in the following 24 hours can be stored in secondary devices. We studied log data from an existing virtual patient record (n=4975 reports) with information on report creation and report first-time visualisation dates, along with contextual information. Bayesian network classifiers were built and compared with logistic regression, revealing high discriminating power (AUC around 90%) and accuracy in predicting whether a report is going to be accessed in the 24 hours after creation.

Keywords: Bayesian networks · Health services · Virtual patient records

1 Introduction

Evidence-based medicine relies on three information sources: patient records, published evidence and the patient itself [25]. Even though great improvements

© Springer International Publishing Switzerland 2015
F. Pereira et al. (Eds.) EPIA 2015, LNAI 9273, pp. 91–102, 2015.
DOI: 10.1007/978-3-319-23485-4_9
and developments have been made over the years, on-demand access to clinical information is still inadequate in many settings, leading to less efficiency as a result of a duplication of effort, excess costs and adverse events [10]. Furthermore, a lot of distinct technological solutions coexist to integrate patient data, using different standards and data architectures which may lead to difficulties in further interoperability [7]. Nonetheless, a lot of patient information is now accessible to health-care professionals at the point of care. But, in some cases, the amount of information is becoming too large to be readily handled by humans or to be efficiently managed by traditional storage algorithms. As more and more patient information is stored, it is very important to efficiently select which one is more likely to be useful [8].

The identification of clinically relevant information should enable an improvement both in user interface design and in data management. However, it is difficult to identify what information is important in daily clinical care, and what is used only occasionally. The main problem addressed here is how to estimate the relevance of health care information in order to anticipate its usefulness at a specific point of care. In particular, we want to estimate the probability of a piece of information being accessed during a certain time interval (e.g. first 24 hours after creation), taking into account the type of data and the context where it was generated and to use this probability to prioritise the information (e.g. assigning clinical reports for secondary storage archiving or primary storage access).

Next section presents background knowledge on electronic access to clinical data (2.1), assessment of clinical data relevance (2.2) and machine learning in health care research (2.3), setting the aim of this work (2.4). Then, section 3 presents our methodology to data processing, model learning, and prediction of within-24h visualisation of clinical data, which results are exposed in section 4. Finally, section 5 finalises the exposition with discussion and future directions.

2 Background

The practice of medicine has been described as being dominated by how well information is collected, processed, retrieved, and communicated [2].

2.1 Electronic Access to Clinical Data

Currently in most hospitals there are great quantities of stored digital data regarding patients, in administrative, clinical, lab or imaging systems. Although it is widely accepted that full access to integrated electronic health records (EHR) and instant access to up-to-date medical knowledge significantly reduces faulty decision making resulting from lack of information [9], there is still very little evidence that life-long EHR improve patient care [4]. Furthermore, there use is often disregarded. For example, studies have indicated that data generated before an emergency visit are accessed often, but by no means in a majority of