Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record

Pedro Pereira Rodrigues[1,2], Cláudia Camila Dias[2], Diana Rocha[2], Isabel Boldt[2],
Armando Teixeira-Pinto[2], Ricardo Cruz-Correia[2]
[1]LIAAD – INESC TEC, Portugal
[2]CINTESIS & Faculty of Medicine of the University of Porto, Portugal
{pprodrigues,camila,isabelboldt,tpinto,rcorreia}@med.up.pt

Abstract

The amount of data currently being produced, stored and used in hospital settings is stressing information technology infrastructure, making clinical reports to be stored in secondary memory devices. The aim of this work was to develop a model that predicts the probability of visualization, within a certain period after production, of each clinical report. We collected log data, from January 2013 till May 2011, from an existing virtual patient record, in a tertiary university hospital in Porto, Portugal, with information on report creation and report first-time visualization dates, along with contextual information. The main factors associated with visualization were defined using logistic regression. These factors were then used as explanatory variables for predicting the probability of a piece of information being accessed after production, using Kaplan-Meier analysis and the Weibull probability distribution. Clinical department, type of encounter and report type were found significantly associated with time-to-visualization and probability of visualization.

1. Introduction

Evidence-based medicine relies on three information sources: patient records, published evidence and the patient itself [1]. 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 [2]. Even though great improvements 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 [3]. Medicine and its practice are deeply dependent on the way information is managed - from how it is recorded and retrieved to how it is actually communicated. This fact led to a strong need of constant improvement of the information management process, what resulted on the development of new information technology solutions and applications in medicine [4]. An important challenge is to guarantee the good working conditions for health professionals to access clinical data while Hospital Information Systems (HIS) are still being developed [5]. Currently, a lot of patient information is accessible to health-care professionals at the point of care. 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.

Implementing a Virtual Patient Record (VPR) system may provide an adequate and cost-effective solution for most clinical information needs. As more and more patient information is stored, it is very important to efficiently select which one is more likely to be useful and promote it in a scenario where scarcity of resources (screen space, storage space, bandwidth and doctors' time) is very real [6]. If we could, for instance, discriminate between reports that will be needed in the next 24 hours from the remaining, we could efficiently decide which ones to store in a faster-accessible memory device. Between May 2003 and May 2004, a virtual patient record (VPR) was designed and implemented at Hospital S. João, a university hospital with over 1350 beds. An agent-based platform Multi-Agent System for Integration of Data (MAID) ensures the communication among various hospital information systems. Clinical reports are retrieved from clinical department information systems (DIS) and stored into a central repository in a browser friendly format. MAID is now running for the last 9 years, regularly scanning 14 DIS and collecting about 7000 new reports each day. Currently, over 340 doctors are using the system on a daily basis.

As the accessibility of biomedical and health-care data with a wide variety of characteristics increases, so does the need to use methods that are able to model the uncertainties inherent to problems and that can actually deal with missing data, enable integrating data from several sources, clearly state statistical dependence and independence, and enable integrating biomedical and clinical background knowledge [7].

The log file is where all actions performed by users of information systems are recorded. Intentionally and

978-1-4799-1053-3/13/$31.00 ©2013 IEEE

originally created and kept for audit purposes, these logs can provide very interesting insights into the information needs of health-care professionals in some particular situations. This is why the study of these logs should not only be carried on to show how the system is used, but also to predict future use of the system and of the data items it contains [8]. A previous study that aimed to determine for how long clinical documents are used by health professionals in a VPR considering the setting of information request and its content, revealed that reports based in different medical encounters have significant differences and these different types of reports have different lifetime [5]. In particular, the usage of past patient information in the VPR described in this study varies significantly according to patient age, type of information, type of hospital encounter and medical cause (main diagnosis) for the encounter. In this work we will focus only on clinical department, encounter type and report type.

In this study we have two objectives: a) to determine the factors associated with the time-to-visualization of a report, and b) using those factors, to develop models that predict the probability of view of each report within 24 hours after production.

2. Methods

We collected usage data from existing virtual patient record (VPR) of Hospital S. João, a university hospital in Porto, Portugal, with information on report creation and report first-time visualization dates, along with contextual information. This study focuses on sessions and report viewings in the VPR from January 2010 to May 2011. The data used in this study was collected with Oracle SQL Developer database system, from the VPR patient database, containing patient's identification and references to the clinical records.

We developed a model with five variables: date of report creation, date of report view, department, encounter type, and report type:

- *date of report creation* - date indicating when the report was created.
- *date of report view* - date indicating when the report was visualized; missing data are treated as non-visualized.
- *department* - identifies the clinical department that generated the report: pneumology, anatomo-pathology, clinical pathology, immunohemotherapy, gastroenterology, cardiothoracic surgery, intensive care, gynaecology endoscopy unit, obstetrics, clinical hematology, paediatric gastroenterology, breast pathology and psychiatry, all modeled dichotomously.
- *encounter type* - identifies the type of encounter that generated the report: block of surgery, consultation, day hospital, inpatient stays, laboratory, radiology and emergency room, all modeled dichotomously.
- *report type* - identifies the specific type of report, different for each department, all modeled dichotomously.

Association between a given report characteristic and the probability of visualization is assessed through logistic regression. The statistical method used to build the predictive model was survival analysis. A Kaplan-Meier [9] analysis allows estimation of survival over time. We estimated the model in R (version 2.11.1) [10] by department, encounter type and report type to analyze the evolution of visualization probability of the report, using the Weibull probability distribution. Association between a given report characteristic and the probability of visualization with a given horizon is assessed through the odds ratio (OR). To assess the quality of the predictive model we address the area between the curve of the Weibull probability distribution model, $\hat{y}(t)$, and the empirical curve computed directly with Kaplan-Meier method,, *y(t)*, i.e. $\int_{t=0}^{T} |\hat{y}(t) - y(t)|\, dt$. As we are dealing with discrete intervals, created by the Kaplan-Meier function, we compute instead $E(t) = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}(i) - y(i)|$, where *n* is the number of reports tested before *t* hours, with *t=24*. This way, we get an estimate of the error for the model when used to predict visualizations within one day.

3. Results

From of 2010 to the first quarter of 2011 the hospital had 534710 records, 210288 (39.33%) from immunohemotherapy, 146209 (27.34%) anatomo-pathology, 127444 (23.83%) clinical pathology, 17333 (3.24%) cardiothoracic surgery, 10355 (1.94%) gastroenterology, 8847 (1.65%) obstetrics, 4873 (0.91%) pneumology, 3988 (0.75%) clinical hematology, 2182 (0.41%) intensive care, 1197 (0.22%) breast pathology and 1110 (0.21%) from the gynaecology endoscopy unit. All departments with less than 1000 reports were not considered for this analysis.

3.1. Relevant report characteristics

Visualization is clearly dependent on type of encounter, with consult reports (OR=0.098) much less likely to be visualized than reports generated during inpatient stays (OR=4.007) or emergency encounters

(OR=5.641). Also, the department related to the produced report is also important to the visualization probability, with the immunohemotherapy (OR=2.418) reports being much more likely to be visualized than, for example, gynecologic endoscopy unit ones (OR=0.106). Given their prevalence (almost 24% of the reports), anatomo-pathology reports need an extra look at, with some types of report being much likely to be visualized than others. Also, although in general gastroenterology reports are only slightly more likely to be visualized (OR=1.018), some particular reports are much more likely to be so (report type 11, OR=6.753), which highlights the need to model report types as well. Other paradigmatic example of this phenomenon is the less visualized group of reports from the department of cardiotoraccic surgery (OR=0.205) where a certain type of report increases the probability of visualization (report type 27, OR=2.762).

Kaplan-Meier survival curves for each studied variable (encounter, department and report type) showed different behaviors across different categories. The main point to stress here is that the studied variables influence not only the probability of visualization, but also the time-to-visualization of a given report, thus supporting the need to create separate survival models for each type of encounter, department or even type of report.

3.2. Survival models for the visualization

Survival models were obtained for each relevant department, type of encounter and report type, and their adjustment to the empirical curves were also inspected (in most cases, the model was well adjusted). For space purposes we only present the curves for two examples (figures 1 and 2). Table 1 presents the 24-hour visualization rates of each time of report. We can see that for this outcome also, the studied variables proved relevant, as different rates were found for each report type and context of report production. The survival models were then used to predict the visualization rate 24 hours after their production; the median error of using those models compared to the curves of actual data, according to the *E(t)* measure, was 6 (min:1, max: 52, for outpatient consults), 17.5 (min=1, max=50, for inpatient stays) and 21 (min=3, max=28, for emergency encounters).

4. Discussion

Our work showed that type of report, type of encounter and department that produced the report, all influence the visualization probability of each particular report. Moreover, the Kaplan-Meier survival analysis showed that it also influences the time-to-visualization of each report. Given this, we have developed separate survival models, based on the Weibull distribution, to predict the visualization of each report, and compared them with the empirical curves given by the Kaplan-Meier analysis. Although not all types of reports can be accurately modeled, we found evidence that the use of these report characteristics can be useful in predictive models, especially to discriminate between reports that will be visualized in the following 24 hours after production, from the remaining ones.

After analyzing the results, we conclude that there is heterogeneity in clinical reports visualization. We had difficulty to build survival models for some types of report, but we identified factors that clearly influence the visualization of reports. Future work is concentrated on: a) implementing a prototype based in these models to calculate the information relevance in real-time; b) evaluating the accuracy and potentially effectiveness of the prototype by analyzing the accuracy of its estimations and changes in user behavior in a real hospital scenario.

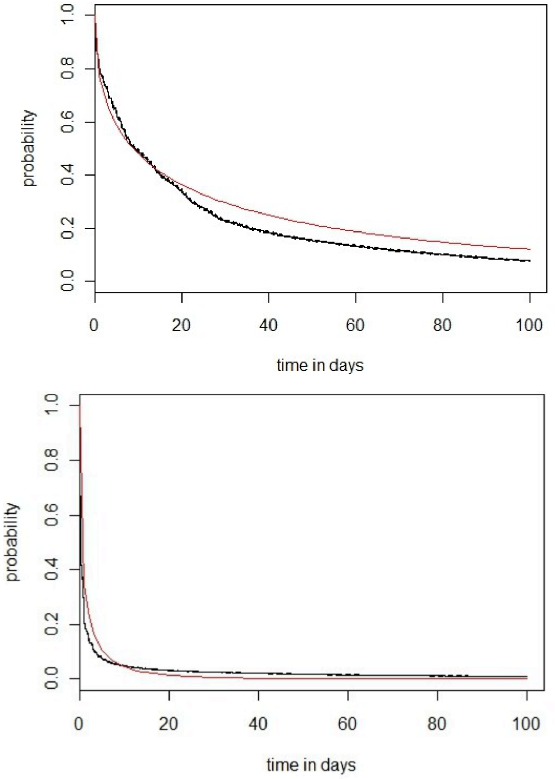

Figure 1: Empirical (black) and predicted (red) survival curves for (top) a type 5 report produced in a outpatient consult by clinical pathology, and (bottom) a type 43 report produced during a inpatient stay by immunohemotherapy.

Table 1: 24-hour visualization rate for different types of report and context of report production.

24-hour visualizations, ***% [95%CI] (n)***	**Outpatient consults**	**Inpatient stays**	**Emergency encounters**
Cardiothoracic surgery			
Report type 3	5 [3.2,7.7] (12)	58 [47.7,65.7] (28)	-
Report type 13	5 [3.9,6.3] (30)	25 [23.4,27.2] (190)	53 [41.7,62.9] (11)
Report type 17	7 [6.6,7.7] (302)	31 [28.5,33.2] (233)	-
Report type 27	-	28 [26.3,30.0] (270)	41 [32.4,48.7] (26)
Report type 29	-	5 [3.2,6.1] (14)	-
Clinical Pathology	20 [19.6,20.3] (5522)	71 [70.6,71.3] (18702)	61 [59.8,61.4] (4988)
Gastroenterology			
Report type 11	53 [41.0,62.4] (11)	88 [85.1,90.9] (73)	-
Report type 15	9 [7.8,10.4] (67)	76 [74.1,77.9] (315)	65 [61.8,67.2] (175)
Report type 16	10 [8.6,11.1] (60)	72 [69.4,73.8] (257)	57 [52.5,61.1] (69)
Clinical Hematology			
Report type 18	-	25 [23.2,26.5] (132)	-
Report type 19	-	23 [20.5,25.2] (61)	-
Pneumology	-	49 [47.3,50.1] (787)	-
Intensive Care	-	43 [41.4,45.5] (313)	-
Anatomo-pathology			
Report type 20	14 [13.7,14.6] (670)	52 [51.6,53.4] (1099)	25 [20.8,29.3] (19)
Report type 21	6 [5.8,7.1] (142)	9 [3.8,13.2] (7)	8 [2.2,14.1] (4)
Report type 22	19 [18.3,19.2] (1254)	36 [35.2,36.1] (3025)	18 [15.7,20.5] (53)
Report type 23	-	23 [19.8,25.5] (28)	-
Report type 24	21 [17.5,24.6] (21)	45 [40.7,48.2] (46)	-
Report type 25	-	34 [31.8,35.9] (121)	-
Report type 26	4 [2.7,5.0] (18)	-	-
Report type 40	23 [21.6,24.4] (153)	38 [30.0,47.4] (318)	39 [30.0,47.4] (12)
Immunohemotherapy	9 [8.9,9.4] (2587)	79 [78.7,79.2] (21226)	89 [88.6 89.2] (8122)
Gynecologic Endoscopy Unit	-	7 [5.0,8.1] (31)	-

Acknowledgements This work was supported by Portuguese Foundation for Science and Technology (FCT) - OPTIM (PTDC/EIA-EIA/099920/2008).

References

[1] Wyatt JC, Wright P: Design should help use of patients' data. Lancet 1998, 352:1375–8.

[2] Cruz-Correia RJ, Vieira-Marques PM, Ferreira AM, Almeida FC, Wyatt JC, Costa-Pereira AM: Reviewing the integration of patient data: how systems are evolving in practice to meet patient needs. BMC medical informatics and decision making 2007, 7:14.

[3] Feied CF, Handler J a, Smith MS, Gillam M, Kanhouwa M, Rothenhaus T, Conover K, Shannon T: Clinical information systems: instant ubiquitous clinical data for error reduction and improved clinical outcomes. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine 2004, 11:1162–9.

[4] Barnett O: Computers in Medicine. JAMA: The Journal of the American Medical Association 1990, 263:2631.

[5] Kuhn KA, Giuse DA, Lapao L, Wurst S: Expanding the Scope of Health Information Systems. Methods of Information in Medicine 2007, 46(4):500-502.

[6] Cruz-Correia RJ, Wyatt JC, Dinis-Ribeiro M, Costa-Pereira A: Determinants of frequency and longevity of hospital encounters' data use. BMC medical informatics and decision making 2010, 10:15.

[7] Lucas P: Bayesian analysis, pattern analysis, and data mining in health care. Current opinion in critical care 2004, 10:399–403.

[8] Rodrigues PP, Dias CC, Cruz-Correia R: Improving clinical record visualization recommendations with bayesian stream learning. In Learning from Medical Data Streams. CEUR-WS.org; 2011, 765:paper4.

[9] Kaplan E. L., Meier P: Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association 1958, 53:457–481.

[10] Team RC: R Language Definition. 2012, 2.