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

Pedro Pereira Rodrigues[1,2,3(✉)], Cristiano Inácio Lemes[4], Cláudia Camila Dias[1,2], and Ricardo Cruz-Correia[1,2]

[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

[2] CIDES-FMUP, Faculty of Medicine of the University of Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal

[3] LIAAD - INESC TEC, Artificial Intelligence and Decision Support Laboratory, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[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

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 times (5% to 20% of the encounters), even when the user was notified of the availability of such data [12].

One usual solution for data integration in hospitals is to consider a virtual patient record (VPR), created by integrating all clinical records, which must collect documents from distributed departmental HIS [3]. Integrated VPR of central hospitals may gather millions of clinical documents, so accessing data becomes an issue. A paradigmatic example of this burden to HIS is the amount of digital data produced in the medical imaging departments, which has increased rapidly in recent years due mainly to a greater use of additional diagnostic procedures, and an increase in the quality of the examinations. The management of information in these systems is usually implemented using Hierarchical Storage Management (HSM) solutions. This type of solution enables the implementation of various layers which use different technologies with different speeds of access, corresponding to different associated costs. However, the solutions which are currently implemented use simple rules for information management, based on variables such as the time elapsed since the last access or the date of creation of information, not taking into account the likely relevance of information in the clinical environment [6].

In a quest to prioritise the data that should be readily available in HIS, several pilot studies have been endured to analyse for how long clinical documents are useful for health professionals in a hospital environment, bearing in mind document content and the context of the information request. Globally, the results show that some clinical reports are still used one year after creation, regardless of the context in which they were created, although significant differences existed in reports created during distinct encounter types [8]. Other results show that half of all visualisations might be of reports more than 2 years-old [20], although this visualisation distribution also varies across clinical department and time of production [21]. Thus, usage of patients past information (data from previous hospital encounters), varied significantly according to the setting of health care and content, and is, therefore, not easy to prioritise.

2.2 Assessment of Clinical Data Relevance

As previously noted, and especially in critical and acute care settings, the age of data is one of the factors often used to assess data relevance, making new information more relevant to the current search. However, studies have shown that some clinical reports are still used after one year regardless of the context in which they were created, although significant differences exist in reports created in distinct encounter types and document content, which contradicts the definition of *old data* used in previous studies. Hence the need to define better rules for recommending documents in encounters.

Classifying the relevance of information based only on the time elapsed since the date of acquisition is clearly inefficient. It is expected that the need to consult an examination at a given time will be dependent on several factors beyond the date of the examination, such as type of examination and the patient's pathology. Thus, a system that uses more factors to identify the relevance of information at a given time would be more efficient in managing the information that is stored in fast memory and slow memory. A recent study from the same group addressed other possibly relevant factors besides document age, including type of encounter (i.e. emergency room, inpatient care, or outpatient consult), department where the report was generated (e.g. gynaecology or internal medicine) and even type of report in each department, but the possibility of modelling visualisations with survival analysis proved to be extremely difficult [21].

Nonetheless, if we could, for instance, discriminate solely between documents 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. Furthermore, we could then rank documents according to their probability of visualisation in order to adjust the graphical user interface of the the VPR, to improve system's usability. By applying regression methods or other modelling techniques it is possible to identify which factors are associated with the usage or relevance of patient data items. These factors and associations can then be used to estimate data relevance in a specific future time interval.

2.3 Machine Learning in Healthcare Research

The definition of clinical decision support systems (most of the times based on expert systems) is currently a major topic since it may help the diagnosis, treatment selection, prognosis of rate of mortality, prognosis of quality of life, etc. They can even be used to administrative tasks like the one addressed by this work. However, the complicated nature of real-world biomedical data has made it necessary to look beyond traditional biostatistics [14] without loosing the necessary formality. For example, naive Bayesian approaches are closely related to logistic regression [22]. Hence, such systems could be implemented applying methods of machine learning [16], since new computational techniques are better at detecting patterns hidden in biomedical data, and can better represent and manipulate uncertainties [22]. In fact, the application of data mining and machine learning techniques to medical knowledge discovery tasks is now a growing research area. These techniques vary widely and are based on data-driven conceptualisations, model-based definitions or on a combination of data-based knowledge with human-expert knowledge [14].

Bayesian approaches have an extreme importance in these problems as they provide a quantitative perspective and have been successfully applied in health care domains [15]. One of their strengths is that Bayesian statistical methods allow taking into account prior knowledge when analysing data, turning the data analysis into a process of updating that prior knowledge with biomedical and health-care evidence [14]. However, only after the 90's we may find evidence of a large interest on these methods, namely on Bayesian networks, which offer a general and versatile approach to capturing and reasoning with uncertainty in medicine and health care [15]. They describe the distribution of probabilities of one set of variables, making possible a two-fold analysis: a qualitative model and a quantitative model, presenting two types of information for each variable.

On a general basis, a Bayesian network represents a joint distribution of one set of variables, specifying the assumption of independence between them, with the inter-dependence between variables being represented by a directed acyclic graph. Each variable is represented by a node in the graph, and is dependent of the set of variables represented by its ascendant nodes; a node X is a ascendant of another node Y if exists a direct arc from X to Y [16]. To give more representational power to the relations represented by the arcs of the graph, it is necessary to associate values to it. The matrix of conditional probability is given for each variable, describing the distribution of probabilities of each variable given its ascendant variables.

After the qualitative and quantitative models are constructed, the next step, and one of the most important, is how to calculate the new probabilities when new evidence is introduced in the network. This process is called inference and works as follows. Each variable has a finite number of categories greater than or equal to two. A node is observed when there is knowledge about the state of that variable. The observed variables have a huge importance because with conditional probabilities they define the prior probabilities of the non observed variables. With the joint probabilities we can calculate the marginal probabilities of each unobserved variable, adding for all categories the probabilities that the variable is in the desired state [15].

2.4 Aim

The aim of this work is the development of a decision support model for discriminating between reports that are going to be useful in the next 24 hours and reports which can be otherwise stored in slower storage devices, since they will not be accessed in the next 24 hours, thus improving performance of the entire virtual patient record system.

3 Data and Methods

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 (see [24] for a description of the system). Clinical documents are retrieved from clinical department information systems (DIS) and stored into a central repository in a browser friendly format. This is done by regularly scanning 14 DIS using different types of agents [17]:

- For each department, a List Agent regularly retrieves report lists from the DIS, with report file references and meta-data, and stores them in the VPR repository.
- The Balancer Agent of that department retrieves the report file references and distributes them to the departmental File Agents.
- File Agents retrieve the actual report files.

As the amount of information available to the agents increases throughout time, there is also an increase in the difficulty of managing that information by humans. Not rarely, a request for a report arrives (after the List Agent has published the existence of that report) before the File Agent was able to retrieve the document. In this cases, an Express Agent is called to retrieve the file, which stresses the entire system's workload, otherwise balanced.

To enable a quantitative analysis (e.g. the likelihood of document access), all actions by users of the VPR are recorded in the log file. Intentionally and 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, although most of the times the quality of these logs is not delivering [5].

3.1 Studied Variables and Outcomes

Data was collected from from the virtual patient record (VPR) with information on report creation and report first-time visualisation dates, along with contextual information. This study focuses on a sample of 5000 reports (2.7% of the entire data for the studied year) and corresponding visualisations, stored in the VPR in 2010. The data used in this study was collected using Oracle SQL Developer from the VPR patient database, containing patient's identification and references to the clinical records. We developed models with seven explanatory variables, including patient data (age and sex), context data (department and type of encounter) and creation time data (hour, day-of-week, daily period), defined as follows. The main outcome of this study was within-24h visualisation of reports.

AgeCat (cat) discretised in decades;
Sex (binary);
Department (cat);
EncType (cat) one of outpatient consult, inpatient care, emergency or other;
Hour (cat) truncated from creation time;
DoW (cat) one of Sun, Mon, Tue, Wed, Thu, Fri or Sat;
Period (cat) one of morning (Hour=7-12), afternoon (13-18), night (19-24) or dawn (1-6);
Visual24h (binary) target outcome, whether the report has been visualised in the first 24 hours after creation or not.

3.2 Model Building and Evaluation

In order to correctly fit the models, only complete cases were considered in the analysis. Logistic regression was applied to all studied variables to predict visualisation. Additionally, two Bayesian network classifiers were built - Naive Bayes (NB) and Tree Augmented Naive Bayes (TAN) - which differ on the number of conditional dependencies (besides the outcome) allowed among variables (NB: zero dependencies; TAN: one dependence), in order to choose the structure which could better represent the problem. Receiver Operating Characteristic (ROC) curve analysis was performed to determine in-sample area under the curve (AUC). Furthermore, to assess the general structure and accuracy of learned models, stratified 10-fold cross-validation was repeated 10 times, estimating accuracy, sensitivity, specificity, precision (positive and negative predictive values) and the area under the ROC curve, for all compared models.

3.3 Software

Logistic regression was done with R package *stats* [18], Bayesian network structure was learned with R package *bnlearn* [23], Bayesian network parameters were fitted with R package *gRain* [11], ROC curves were computed with R package *pROC* [19], and odds ratios (OR) were computed with R package *epitools* [1].

4 Results

A total of 4975 reports were included in the analysis. The main characteristics of the reports are shown in Table 1, which were generated from patients with a mean (std dev) age of 55.5 (20.5). Less than 23% of the reports were visualised in the 24 hours following their creation, which were nonetheless more from female patients (almost 55%) with a 24h-visualisation OR=1.51 (95%CI [1.32,1.72]) for female-patient reports. Also significant was the context of report creation, with more reports being created in inpatient care (44.4%) and outpatient consults (41.4%), although compared with the latter context, 24-hour visualisations are more likely for reports generated in inpatient care (OR=8.60 [7.04,10.59]) or in the emergency room (OR=14.50 [11.22,18.83]). Regarding creation time, morning (OR=1.22 [1.05,1.41]), night (OR=1.82 [1.46,2.28]) and dawn (OR=2.88 [2.03,4.07]) have all higher 24-hour visualisation likelihood than the afternoon period.

4.1 Qualitative Analysis of the Bayesian Network Model

Figure 1 presents the qualitative model for the Tree-Augmented Naive Bayes network, where interesting connections can be extracted from the resulting model. First, patient's data features are associated. Then, creation time data and context data are also strongly related. However, the most interesting feature is probably the department that created the report, since this was chosen by the algorithm as ancestor of patient's age, time of report creation and type of encounter.

4.2 In-Sample Quantitative Analysis

For a quantitative analysis, Figure 2 presents the in-sample ROC curves for logistic regression (left), Naive Bayes (centre) and TAN (right). As expected, increasing model complexity enhances the in-sample AUC (LR 88.6%, NB 86.9% and TAN 90.7) but, globally, all models presented good discriminating power towards the outcome.

Table 1. Basic characteristics of included reports: patient's data (sex and age), report creation context (department, encounter) and time (day of week, daily period) data.

	Visualised in 24 hours		
	No	**Yes**	**Total**
Outcome, *n (%)*	3846 (77.3)	1129 (22.7)	4975 (100)
Female, *n (%)*	1716 (44.6)	619 (54.8)	2335 (46.9)
Age, $\mu(\sigma)$	54.6 (19.8)	58.5 (22.4)	55.5 (20.5)
AgeCat, *n (%)*			
[0,10[	97 (2.5)	59 (5.2)	156 (3.1)
[10,20[	58 (1.5)	23 (2.0)	81 (1.6)
[20,30[	215 (5.6)	40 (3.5)	255 (5.1)
[30,40[	583 (15.2)	115 (10.2)	698 (14.0)
[40,50[	597 (15.5)	122 (10.8)	719 (14.5)
[50,60[	601 (15.6)	150 (13.3)	751 (15.1)
[60,70[	710 (18.5)	199 (17.6)	909 (18.2)
[70,80[	554 (14.4)	207 (18.3)	761 (15.3)
[80,90[	372 (9.67)	181 (16.0)	553 (11.1)
[90,100[	55 (1.4)	31 (2.8)	86 (1.7)
$\geq$100	4 (0.1)	2 (0.2)	6 (0.1)
Encounter Type, *n (%)*			
Outpatient consult	1940 (50.4)	120 (10.6)	2060 (41.4)
Inpatient care	1442 (37.5)	768 (68.0)	2210 (44.4)
Emergency room	217 (5.6)	19 (1.7)	236 (4.7)
Other	247 (6.4)	222 (19.7)	469 (9.4)
Department, *n (%)*			
1	76 (2.0)	11 (1.0)	87 (1.8)
2	1626 (42.3)	55 (4.9)	1681 (33.8)
3	646 (16.8)	469 (41.5)	1115 (22.4)
5	1057 (27.5)	529 (46.9)	1586 (31.9)
6	154 (4.0)	23 (2.0)	177 (3.6)
7	89 (2.3)	22 (2.0)	111 (2.2)
9	11 (0.3)	7 (0.6)	18 (0.4)
10	10 (0.3)	1 (0.1)	11 (0.2)
12	139 (3.6)	11 (1.0)	150 (3.0)
13	23 (0.6)	0 (0)	23 (0.3)
16	5 (0.1)	0 (0)	5 (0.1)
21	10 (0.3)	1 (0.1)	11 (0.2)
Day-of-Week, *n (%)*			
Mon	728 (18.9)	303 (26.8)	1031 (20.7)
Tue	671 (17.5)	291 (25.8)	962 (19.3)
Wed	743 (19.3)	208 (18.4)	951 (19.1)
Thu	804 (20.9)	35 (3.1)	839 (16.9)
Fri	673 (17.5)	92 (8.2)	765 (15.4)
Sat	122 (3.2)	99 (8.7)	221 (4.4)
Sun	105 (2.7)	101 (9.0)	206 (4.1)
Daily Period, *n (%)*			
Morning	1768 (46.0)	521 (46.2)	2289 (46.0)
Afternoon	1661 (43.2)	402 (35.6)	2063 (41.5)
Night	331 (8.6)	146 (13.0)	477 (9.6)
Dawn	86 (2.2)	60 (5.3)	146 (2.9)

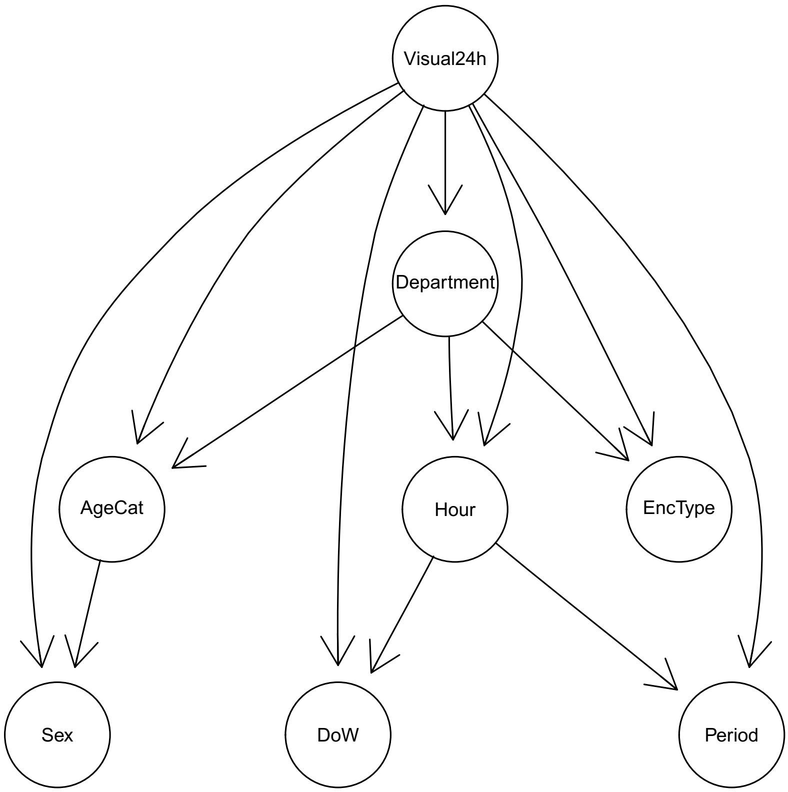

Fig. 1. Tree-Augmented Naive Bayes for predicting within 24h visualisation of clinical reports in the virtual patient record.

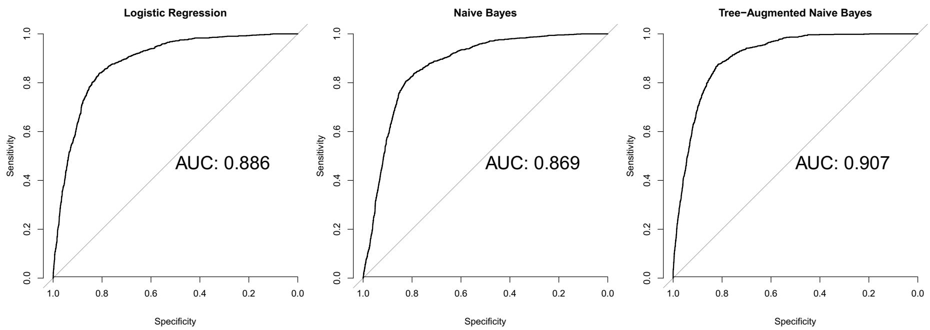

Fig. 2. In-sample ROC curves for logistic regression (left), naive Bayes (centre) and Tree-Augmented Naive Bayes (right).

Table 2. Validity assessment averaged from 10 times stratified 10-fold cross-validation for logistic regression (LR), naive Bayes (NB) and Tree-Augmented Naive Bayes (TAN).

Measure, % [95%CI]	**LR**	**NB**	**TAN**
Accuracy	82.30 [82.08,82.52]	82.43 [82.14,82.72]	82.80 [82.51,83.09]
Sensitivity	41.33 [40.50,42.16]	60.68 [59.81,61.55]	64.12 [63.36,64.89]
Specificity	94.33 [94.07,94.58]	88.81 [88.50,89.12]	88.28 [87.96,88.61]
Precision (PPV)	68.40 [67.45,69.35]	61.53 [60.80,62.25]	61.75 [61.04,62.47]
Precision (NPV)	84.57 [84.39,84.76]	88.51 [88.29,88.74]	89.35 [89.15,89.56]
AUC	87.58 [87.27,87.89]	86.37 [86.04,86.70]	85.50 [85.13,85.88]

4.3 Bayesian Network Generalisable Cross-Validation

In order to assess the ability of the models to generalise beyond the derivation cohort, cross-validation was endured. Table 2 presents the result of the 10-times-repeated stratified 10-fold cross-validation. Although the more complicated model loses in terms of AUC (85% vs 87%), it brings advantages to the precise problem of identifying reports that should be stored in secondary memory as they are less likely to be visualised in the next 24 hours, since it reveals a negative precision of 89% vs 88% (NB) and 84% (LR). Along with this result, it is much better at identifying reports that are going to be needed, as sensitivity rises from 41% (LR) to 64%. Future work should consider different threshold values for the decision boundary (here, 50%) in order to better suit the model to the sensitivity-specificity goals of the problem a hands.

5 Concluding Remarks and Future Work

The main contribution of this work is the preliminary study for the development of a decision support model for discriminating between reports that are going to be useful in the next 24 hours and reports which can be safely stored in secondary memory, since they will not be accessed in the next 24 hours.

An initial sample of clinical reports was used to derive Bayesian network models which were then compared with a logistic regression model in terms of in-sample discriminating power and generalisable validity with cross-validation. The studied data was in accordance with previous works in terms of the relevance that some factors may have on the likelihood of visualisation of clinical reports, e.g. department and type of encounter that produced the report [21]. Additionally, patient data and time of report creation were also found relevant for the global model of predicting within 24-hour visualisations.

Given that the main objective of this project is to enable a clear decision on whether a report can safely be stored in secondary memory or not, focus should be given to negative precision, since it represents the probability that a report marked by the system to be stored away is, in fact, irrelevant for the present day. The Bayesian network models achieved negative precision of around 89%, while keeping specificity high (also around 88%).

Future work will be concentrated in a) exploring other variables that might influence the likelihood of visualisation of clinical reports (e.g. actual data from the report, patient's diagnosis, etc.); b) exploiting the maximum amount of data from the log file of the virtual patient record (e.g. 2010 comprises of more than 184K reports); and c) inspecting the usefulness of temporal Bayesian network models [13] for the precise problem of relevance estimation.

Overall, this study presents Bayesian network models as useful techniques to integrate in a virtual patient record that needs to prioritise the accessible documents, both in terms of user-interface optimisation and data management procedures.

Acknowledgments. The authors acknowledge the help of José Hilário Almeida during the data gathering process.

References

1. Aragon, T.J.: epitools: Epidemiology Tools (2012)
2. Barnett, O.: Computers in medicine. JAMA: the Journal of the American Medical Association **263**(19), 2631 (1990)
3. Bloice, M.D., Simonic, K.M., Holzinger, A.: On the usage of health records for the design of virtual patients: a systematic review. BMC Medical Informatics and Decision Making **13**(1), 103 (2013)
4. Clamp, S., Keen, J.: Electronic health records: Is the evidence base any use? Medical Informatics and the Internet in Medicine **32**(1), 5–10 (2007)
5. Cruz-Correia, R., Boldt, I., Lapão, L., Santos-Pereira, C., Rodrigues, P.P., Ferreira, A.M., Freitas, A.: Analysis of the quality of hospital information systems audit trails. BMC Medical Informatics and Decision Making **13**(1), 84 (2013)
6. Cruz-Correia, R., Rodrigues, P.P., Freitas, A., Almeida, F., Chen, R., Costa-Pereira, A.: Data quality and integration issues in electronic health records. In: Hristidis, V. (ed.) Information Discovery on Electronic Health Records, chap. 4. Data Mining and Knowledge Discovery Series, pp. 55–95. CRC Press (2009)
7. Cruz-Correia, R.J., Vieira-Marques, P.M., Ferreira, A.M., Almeida, F.C., Wyatt, J.C., Costa-Pereira, A.M.: Reviewing the integration of patient data: how systems are evolving in practice to meet patient needs. BMC Medical Informatics and Decision Making **7**, 14 (2007)
8. Cruz-Correia, R.J., Wyatt, J.C., Dinis-Ribeiro, M., Costa-Pereira, A.: Determinants of frequency and longevity of hospital encounters' data use. BMC Medical Informatics and Decision Making **10**, 15 (2010)
9. Dick, R., Steen, E.: The Computer-based Patient Record: An Essential Technology for HealthCare. National Academy Press (1997)
10. Feied, C.F., Handler, J.A., Smith, M.S., 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 **11**(11), 1162–1169 (2004)
11. Højsgaard, S.: Graphical independence networks with the gRain package for R. Journal of Statistical Software **46**(10) (2012)
12. Hripcsak, G., Sengupta, S., Wilcox, A., Green, R.: Emergency department access to a longitudinal medical record. Journal of the American Medical Informatics Association **14**(2), 235–238 (2007)

13. Lappenschaar, M., Hommersom, A., Lucas, P.J.F., Lagro, J., Visscher, S., Korevaar, J.C., Schellevis, F.G.: Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity. Journal of Clinical Epidemiology **66**, 1405–1416 (2013)
14. Lucas, P.: Bayesian analysis, pattern analysis, and data mining in health care. Current Opinion in Critical Care **10**(5), 399–403 (2004)
15. Lucas, P.J.F., van der Gaag, L.C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine **30**(3), 201–214 (2004)
16. Mitchell, T.M.: Machine Learning. McGraw-Hill (1997)
17. Patriarca-Almeida, J.H., Santos, B., Cruz-Correia, R.: Using a clinical document importance estimator to optimize an agent-based clinical report retrieval system. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pp. 469–472 (2013)
18. R Core Team: R: A Language and Environment for Statistical Computing (2013)
19. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics **12**, 77 (2011)
20. Rodrigues, P.P., Dias, C.C., Cruz-Correia, R.: Improving clinical record visualization recommendations with bayesian stream learning. In: Learning from Medical Data Streams, vol. 765, p. paper4. CEUR-WS.org (2011)
21. Rodrigues, P.P., Dias, C.C., Rocha, D., Boldt, I., Teixeira-Pinto, A., Cruz-Correia, R.: Predicting visualization of hospital clinical reports using survival analysis of access logs from a virtual patient record. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, pp. 461–464 (2013)
22. Schurink, C.A.M., Lucas, P.J.F., Hoepelman, I.M., Bonten, M.J.M.: Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. The Lancet infectious diseases **5**(5), 305–312 (2005)
23. Scutari, M.: Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software **35**, 22 (2010)
24. Vieira-Marques, P.M., Cruz-Correia, R.J., Robles, S., Cucurull, J., Navarro, G., Marti, R.: Secure integration of distributed medical data using mobile agents. Intelligent Systems **21**(6), 47–54 (2006)
25. Wyatt, J.C., Wright, P.: Design should help use of patients' data. Lancet **352**(9137), 1375–1378 (1998)