Measurement 93 (2016) 480–489

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Design and evaluation of a decision support system for pain management based on data imputation and statistical models

Nuno Pombo [a,b,*], Paulo Rebelo [c], Pedro Araújo [a,b], Joaquim Viana [d]

[a] Instituto de Telecomunicações, Universidade de Beira Interior, Covilhã, Portugal
[b] Department of Informatics, Universidade de Beira Interior, Covilhã, Portugal
[c] Department of Mathematics, Universidade de Beira Interior, Covilhã, Portugal
[d] Faculty of Health Sciences, Universidade de Beira Interior, Covilhã, Portugal

ARTICLE INFO

Article history:
Received 6 July 2014
Received in revised form 27 March 2016
Accepted 3 July 2016
Available online 5 July 2016

Keywords:
Clinical decision support system
Pain assessment
Data imputation
Linear regression
Analysis of variance

ABSTRACT

The self-reporting of pain complaints is considered the most accurate pain assessment method and represents a valuable source of data to computerised clinical decision support systems (CCDSS) for pain management. However, the subjectivity and variability of pain conditions, combined with missing data, are constraints on the usefulness and accuracy of CCDSS. Based on data imputation principles, together with several mathematical models, this paper presents a CCDSS, the Patient Oriented Method of Pain Evaluation System (POMPES), that produces tailored alarms, reports, and clinical guidance based on collected patient-reported data. This system was tested using clinical data collected during a six-week randomised controlled trial involving thirty-two volunteers recruited from an ambulatory surgery department. The decisions resulting from the POMPES were fully accurate when compared with clinical advice, which proves the ability of the system to cope with missing data and detect either stability or changes in the self-reporting of pain.

1. Introduction

In recent years, computerised clinical decision support systems (CCDSS) have largely been used to enhance health by providing health care professionals (HCP) and patients with knowledge and individualised information that is intelligently selected or presented at appropriate times. These systems may lead to better clinical guidance, patient perspectives on their condition, and HCP practices [1–3], which are established based on decisions made not only on the basis of their perception and experience but also supported by the collected data. In addition, automated alerts, reminders, and the availability of information when and where it is needed are features intended to optimise the clinical workflow [4,5] and thus improve the quality of treatment. When this occurs, the computerised system supports clinical decisions instead of acting as merely stand-alone software operating in parallel to HCP. Thus, designing CCDSS models to represent medical concepts and tasks, such as diagnosis, treatment, and screening, poses several challenges when the goal is to produce systems with the capability to make better use of the existing data and to extend the information on which decisions are based. Moreover, the problem of missing values commonly arises in the collected data [6,7] processed by the CCDSS, which may lead to incorrect and inaccurate analyses.

In this context, mathematical models are being increasingly adopted by the CCDSS with the aim of enhancing data analysis and processing to produce patient-oriented recommendations that are delivered to HCP [8–10]. Furthermore, several techniques of data imputation have been developed to compensate for missing data [11] with the aim of producing more precise and reliable systems. These improvements related to CCDSS are even more significant when these systems are applied to manage patient-specific conditions with large variability and more difficult assessment, such as pain symptoms. In fact, the subjectivity of pain relies on physiological, neurological, and psychological aspects representing a multidimensional experience [12–14] that raises several challenges to the definition of correct treatments [15]. In addition, because the self-reporting of pain complaints is considered the most accurate pain assessment method [16–18], these data are of particular importance to the reliability of CCDSS applied to pain management, and therefore it is critical to solve the issue of gaps in the dataset.

The aim of this study is to present and validate a CCDSS, the Patient Oriented Method of Pain Evaluation System (POMPES),

* Corresponding author at: Department of Informatics, Universidade of Beira Interior, Covilhã, Portugal.

E-mail addresses: ngpombo@ubi.pt (N. Pombo), rebelo@ubi.pt (P. Rebelo), paraujo@di.ubi.pt (P. Araújo), jsviana@fcsaude.ubi.pt (J. Viana).

http://dx.doi.org/10.1016/j.measurement.2016.07.009
0263-2241/

which comprises data imputation principles and adaptable mathematical models for the production of tailored alarms, reports, and clinical guidance based on collected patient-reported data. The paper is organised as follows. Section 2 presents the background and state-of-the-art research with a focus on the data imputation techniques and algorithms used by CCDSS, and Section 3 addresses the monitoring system in which the proposed decision model was applied and tested. Section 4 presents a detailed explanation of the mathematical concepts behind the system, and the results are presented in Section 5. Finally, Section 6 concludes the paper.

2. Background

The literature comprises a large variety of algorithms used by CCDSS applied to pain measurement. These algorithms are summarised below. First, rule-based algorithms [19–26] have been used, including decision tree algorithms, such as ID3 [27], C4.5 [28], and algorithms for optimising and/or ranking of decision rules and variables, namely CN2 [29], CART [30], ITRULE [31] and ILLM [32]. Rule-based algorithms produce understandable classifications, but some limitations are present, such as the overspecialisation or the inability to learn from incomplete data [33–35].

Second, artificial neural networks (ANNs) [36–40], composed either of single-layer or multi-layer perceptrons, generate an output set where each element represents a particular classification for the input set. This is achieved via the propagation of estimated weights through the nodes of the network obtained from a batch of training in a repeated manner. ANNs are robust if given noisy data and can represent complex functions [41,42], whereas the inability to explain decision, present data clearly [34,43], and determine the adequate size of the hidden layer (when multiple layers are used) are observed disadvantages [44,45].

Third, rough and fuzzy sets [46–51] have also been used and encompass rough set [52] and fuzzy set [53] models. The rough set is obtained from the difference between two sets of elements: those that certainly belong to the set and those that most likely belong to the set. This classification approach may work with continuous or discrete data, but the algorithms required to extract knowledge (by creating reducts and rules) are computationally expensive and may cause problems for large datasets [54,55]. The fuzzy set represents a probabilistic logic model that uses reasoning to explain whether an event is about to happen, which means that every element within the set has a degree of relevance (a.k.a. membership) varying between 0 (or false) and 1 (or true). Thus, fuzzy sets are suitable for representing uncertain or flexible information [56], despite difficulties in estimating membership functions [57]. Finally, there are other methods [58–68], such as Bayes' theorem (a.k.a. Bayes' rule) [69], naive Bayes [70], Bayesian networks [71], logistic regression (LR) [72], and support vector machine (SVM) [73]. Bayesian algorithms are time-consuming models and require a thorough knowledge of parameters [74,75]. LR is less susceptible to overfitting [76] but is unsuitable for dealing with non-linear problems [77]. SVM has good generalisation ability, but uses a formalism that is often unsuitable for interpretation by human experts [78]. Furthermore, several authors have presented a variety of models that may also be considered as they are largely used for comparison of the collected data, estimating treatment effects, assessing outcomes and, consequently, to determine the accuracy and validity of computerised systems applied to pain measurement. These models differ from Fisher's test [79,80], Pearson's test [79,81–83], and the *t*-test [80,84–89] methods that are based on the analysis of variance and covariance such as ANOVA [83,90–96], ANCOVA [97–99], MANOVA [87,94,100], and MANCOVA [101].

Regardless the selected algorithm, the design of CCDSS for pain management faces an additional challenge related to missing data. In this study, the existing techniques to address missing data were categorised into the following categories:

- Deletion methods [102,103]: These involve either discarding all records with missing values for at least one variable (listwise deletion) or discarding only instances with missing values for the less important variables (pairwise deletion). Simplicity is the main advantage, whereas the reduction of the statistical power and inability to perform comparison analysis (when pairwise deletion is used) are limitations.
- Supplement methods [104–108]: These involve replacing missing data with computed values estimators (e.g., mean, median, mode and hot-deck) or applying regression imputation methods such as linear, multiple linear and logistic regression. The hot-deck imputation estimates missing values in incomplete records using values from similar complete records. The adoption of imputation estimators based on mean, median or mode is likely to reduce the variability of data. Moreover, mean imputation is affected by the presence of outliers, and, for that reason, in some cases the median imputation is more appropriate and may create spikes in the data distribution. Regression imputation replaces missing data based on cases with complete data. This technique may reduce the problem of spikes, but it may overestimate the model fit and weaken the variance.
- Model-based methods [105,109–113]: These involve replacing missing data with more sophisticated models, such as maximum likelihood, multiple imputation and machine learning techniques such as SVM or ANN. Maximum likelihood methods estimate the missing data using a set of records that is most likely to have resulted in the observed data. Multiple imputation uses a model to replace missing data multiple times. The main difficulty lies in designing a suitable method to perform the imputation [114] (Monte Carlo Markov Chain and Multiple Imputation by Chained Equations are often used). Maximum likelihood and multiple imputation may produce unbiased estimates. Nearest-neighbour imputation determines the similarity of two records using the distance between them. This method can handle records with multiple missing values and takes into account the correlational structure of data [115]. However, the method is time consuming, and requires the choice of distance function.

Thus, our approach is based on data imputation principles, together with several mathematical models to determine either stability or change in pain intensity obtained from self-reporting.

3. Monitoring system

The proposed CCDSS aims to support HCP during the monitoring of patients suffering with pain, independently of their conditions, and self-reporting frequently. This self-reporting is validated using a computerised pain monitoring system [116] developed by our research team. As shown in Fig. 1, the proposed system runs server-side and is integrated with a Personal Health Record (PHR) accessible to HCP and patients. The input set of this system is based on patients' self-reported data inserted directly into the PHR using a browser or collected via mobile device and sent to the PHR using web services (WS). Finally, the monitoring software combines the outcome provided by the CCDSS with the patients' monitoring rules (e.g., value-oriented messages as presented with more detail in the next section) defined in the PHR to send alarms and alerts messages to either HCP or patients.