Computers in Biology and Medicine 82 (2017) 119–129

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Comprehensible knowledge model creation for cancer treatment decision making

Muhammad Afzal[a], Maqbool Hussain[a], Wajahat Ali Khan[a], Taqdir Ali[a], Sungyoung Lee[a,*], Eui-Nam Huh[a], Hafiz Farooq Ahmad[b], Arif Jamshed[c], Hassan Iqbal[d], Muhammad Irfan[c], Manzar Abbas Hydari[c]

[a] *Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, South Korea*
[b] *College of Computer Sciences and Information Technology (CCSIT), King Faisal University, Alahsa, Saudi Arabia*
[c] *Shaukat Khanum Memorial Cancer Hospital and Research Center, Lahore, Pakistan*
[d] *Department of Otolaryngology and Head and Neck Surgery, The Ohio State University, USA*

ARTICLE INFO

Keywords:
Knowledge acquisition
Algorithm selection
Prediction model
Decision support
Education support

ABSTRACT

Background: A wealth of clinical data exists in clinical documents in the form of electronic health records (EHRs). This data can be used for developing knowledge-based recommendation systems that can assist clinicians in clinical decision making and education. One of the big hurdles in developing such systems is the lack of automated mechanisms for knowledge acquisition to enable and educate clinicians in informed decision making. *Materials and Methods*: An automated knowledge acquisition methodology with a comprehensible knowledge model for cancer treatment (CKM-CT) is proposed. With the CKM-CT, clinical data are acquired automatically from documents. Quality of data is ensured by correcting errors and transforming various formats into a standard data format. Data preprocessing involves dimensionality reduction and missing value imputation. Predictive algorithm selection is performed on the basis of the ranking score of the weighted sum model. The knowledge builder prepares knowledge for knowledge-based services: clinical decisions and education support. *Results*: Data is acquired from 13,788 head and neck cancer (HNC) documents for 3447 patients, including 1526 patients of the oral cavity site. In the data quality task, 160 staging values are corrected. In the preprocessing task, 20 attributes and 106 records are eliminated from the dataset. The Classification and Regression Trees (CRT) algorithm is selected and provides 69.0% classification accuracy in predicting HNC treatment plans, consisting of 11 decision paths that yield 11 decision rules. *Conclusion*: Our proposed methodology, CKM-CT, is helpful to find hidden knowledge in clinical documents. In CKM-CT, the prediction models are developed to assist and educate clinicians for informed decision making. The proposed methodology is generalizable to apply to data of other domains such as breast cancer with a similar objective to assist clinicians in decision making and education.

1. Introduction

Cancer is a major public health problem worldwide and is currently the cause of 1 in 4 deaths in the United States [1], making it the second leading cause of death in the US [2]. In a very recent review, it is stated that more than 1 in 3 people in the United Kingdom will develop some form of cancer during their lifetime [3]. It is also one of the most complex chronic diseases, requiring a guideline- and protocol-driven team-based approach to care [4]. Management of treatment plans and operational inefficiencies greatly influences the safety, quality, efficacy, and cost of care [5]. The authors in [6] mentioned that health systems can influence cancer outcomes through three mechanisms: coverage, innovation, and quality of care. Among these, computerized systems can greatly help to improve quality of care by reducing the chance of errors and time. Most cancer care systems are developed in a group setting based on the requirements established by a health provider organization. For future analysis, the clinical data are either manually analyzed or entered into a computer system by humans. However, manual methods generate unintentional errors [7], and deliberate modification of data may influence the quality of the information [8].

With increasing use of information technology and wider adoption of electronic health records (EHRs), there is a need to expand the use of

* Corresponding author.
E-mail addresses: muhammad.afzal@oslab.khu.ac.kr, muhammad.afzal@oslab.khu.ac.kr (M. Afzal), maqbool.hussain@oslab.khu.ac.kr (M. Hussain), wajahat.alikhan@oslab.khu.ac.kr (W. Ali Khan), taqdir.ali@oslab.khu.ac.kr (T. Ali), sylee@oslab.khu.ac.kr (S. Lee), drhassaniqbal@gmail.com, iqbal.56@osu.edu (Hassan Iqbal),.

http://dx.doi.org/10.1016/j.compbiomed.2017.01.010
Received 4 July 2016; Received in revised form 17 January 2017; Accepted 17 January 2017

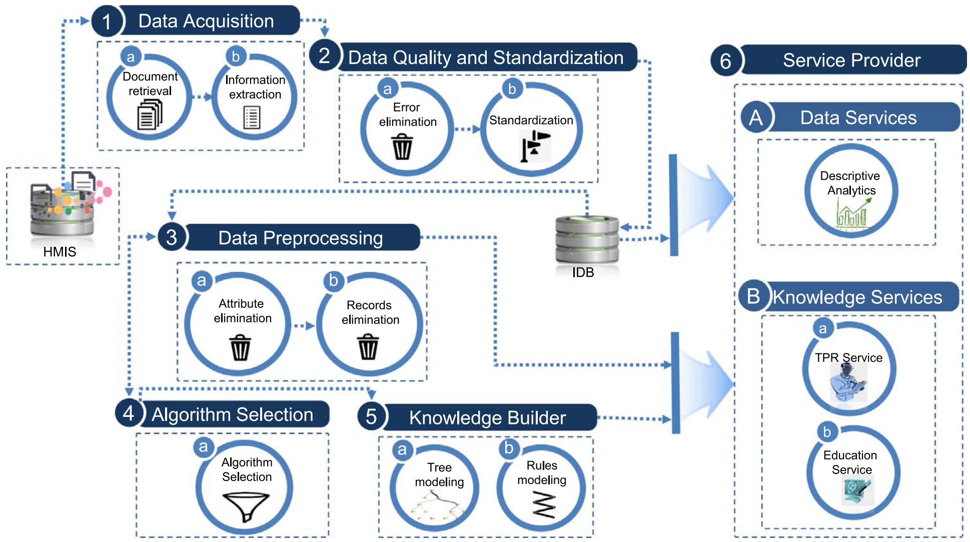

Fig. 1. Functional workflow diagram of the proposed methodology. HMIS=Health management information system, TPR=Treatment plan recommendation, IDB=Intermediate database.

clinical data to support clinical decisions and research [9,10]. EHRs have transformed the way healthcare is carried out [10] and have increased the role and acceptance of clinical decision support systems (CDSSs) in daily clinical practice [11]. It is unrealistic to expect the data retrieved from an EHR to be 100% complete and error free [12]; therefore, it is necessary to address data quality. Similarly, quite often in clinical data, the label attributes and values are not consistent in terminology. Lack of standardized data and terminology hinders the ability to mine the data for patterns and patient outcomes [12]. A major barrier to achieving the maximum benefit from these opportunities is the large amount of valuable clinical knowledge buried within clinical narratives in patient records [9,13,14] exists in raw data form. Raw data simply exists and has no significance beyond its existence [15], whereas knowledge is useful because of its explicit and decision-oriented nature.

One of the important applications in data mining is the use of statistical approaches for knowledge mining from extracted information in order to create predictive models [10]. The predictive models are used for finding patterns in the data to help in the diagnosis and treatment of current and future patients [10,16]. However, these models require well-prepared, correct, and structured data prior to their application in a domain [17]. Preprocessing and selection of an appropriate machine learning algorithm are necessary and challenging tasks that need to be addressed prior to building knowledge for recommendations. In preprocessing, one frequently occurring issue is the missing values in data which requires resolution with different value imputation techniques. Missing value imputation exploits information about the data to estimate the missing entries [18]; and this is a common problem in statistical analysis [19] and in data mining approaches [12]. It occurs in almost all medical and epidemiological research [20]. Moreover, the role and nature of the appropriate machine learning algorithm are inevitable to consider for a particular problem to achieve a target objective. As applications are a necessary precondition for the success of machine learning [21], a machine learning algorithm requires alignment with a target application.

There are efforts made in the area of developing and evaluating decision support system and services for cancer patient care [22–26]. For instance, researchers of work in [22] aimed at creating an information technology oriented decision support system for breast cancer treatment based on data mining techniques and clinical practice guidelines. For head and neck cancer treatment, authors of [23] introduced a three phase knowledge acquisition and validation model that uses data-driven approach for initial level knowledge acquisition which in turn validated using clinical practice guidelines. A study in [24] aimed to develop and assess the CDSS feasibility for breast cancer (BC) treatment planning based on clinical practice guidelines, which they reported that the initial application achieved very encouraging results. A large population-based data set is collected in study [26] to create a clinical decision support system (CDSS) for colon cancer (CC) patients to identify the real-time overall survival using Bayesian Belief Network Model. Based on our analysis, the majority of these systems lacks the connection of a clinical decision support system that is developed on the basis of clinical data extracted from structured/unstructured clinical documents of patients registered in the hospital management system.

In this paper, we propose an automated knowledge acquisition methodology with a comprehensible knowledge model, called CKM-CT. This methodology recommends and predicts the appropriate treatment plan for head and neck cancer (HNC) patients based on the information retrieved from the clinical documents. The proposed methodology involves a set of key functions: data acquisition that acquires clinical data from the clinical documents; data quality and standardization that verifies the quality of data by correcting erroneous data and transforms the data variations into a standard form; data preprocessing that

Table 1
Document retrieval service specifications.

Clinical note/report	Input specification	Method type
Patient Note Information contains structured and semi-structured data	•frmMrno [Patient medical record number] •frmNotesType [Note type: F] •frmNotesFromDate [start date] •frmNotesToDate, [end date]	GET [JSON]
Patient Drug Information contains semi-structured data	•frmMrno [Patient medical record number] •frmDrugFromDate [Drug start date] •frmDrugToDate	GET [JSON]
Patient Histopathology Information contains unstructured data	•frmMrno [Patient medical record number] •frmStartDate [report start date] •frmEdDate [report end date]	GET [JSON]
Patient Chemo Treatment Summary Information contains semi-structured and unstructured data	•frmMrno [Patient medical record number] •frmFromDate [report start date] •frmToDate, [report end date]	GET [JSON]