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,e], Maqbool Hussain[a,e], 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
[e] Department of Software, Sejong University, South Korea

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 clinical data to support clinical decisions and research [9,10]. EHRs have

* 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

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 reduces the dimensionality of the data by selecting the relevant and domain-significant attributes and handles the missing values; algorithm selection that selects the most appropriate algorithm from the candidate machine learning algorithms; and a knowledge builder that builds the knowledge for knowledge-based services: clinical decision support and education support.

2. Motivation

A wealth of information can be found in the form of clinical documents, notes, and reports, which is contributing to day-to-day clinical practice. Unfortunately, due to the lack of specialized systems for automatically information extraction, important clinical data are either underutilized or poorly managed through the intensive involvement of data entry operators. Typically, resident doctors perform the task of data entry, which can be properly managed in clinical practice, if the task is automatically resolved; this process affects the performance of clinicians in terms of diversion from clinical to non-clinical activities. Based on our analysis, discussions, and interviews with oncologists, we noticed several inadequacies such as data entry mistakes that ultimately reduce the overall quality of clinical practice. Literature shows that the principal source of mistakes when entering data is user error [27]. During manual data entry, there is a risk of entering incorrect clinical values, which can affect all subsequent analysis steps performed on this data. In addition, the time spent on manual entries is valuable and would be regained through the use of automated computerized methods and programs. Our main motivation of the proposed methodology is to minimize the effort clinicians spend on manual data entries and assist them in clinical decisions and research. Moreover, the automatic methods are reusable for other types of cancer, as many of the concepts are common. For example, clinical staging is a common concept used for different types of cancer. Any sort of automation, whether it is at the time of data acquisition or data preparation for statistical methods, can become a reusable component of the system.

3. Materials and methods

To fulfill the target objectives, we present the functional workflow of the proposed CKM-CT methodology in Fig. 1. The methodology involves five core functions: data acquisition, data quality assessment and language standardization, data preprocessing, algorithm selection, and a knowledge builder. On the basis of these five functions, the service provider provides data and knowledge services including descriptive analytics, clinical decision support, and education support.

3.1. Data acquisition

The data acquisition function is designed to acquire data from the source system, the health management information system (HMIS). Interaction with the HMIS system happens through a web service and the requested documents are retrieved and then passed to the information extraction function for extraction of the desired information.

3.1.1. Document retrieval

The document retrieval function acquires the data from the source system through the RESTful web service using the JSON format [28]. It retrieves four types of documents from HMIS: patient notes, drug reports, histopathology reports, and chemo treatment summary reports. Details on the specification of each RESTful service for related documents are provided in Table 1.

The information extraction function extracts information from the retrieved documents on the basis of attribute mapping. The RESTful service provides information in a key-value format that is mapped to the name-value format of a column in a relational database called intermediate database (IDB in Fig. 1). The information in the IDB is categorized into two groups: pre-treatment attributes and post-treatment attributes, as shown in Table 2. The 22 pre-treatment attributes are categorized as demographics, risk factors, diagnosis, clinical staging, treatment, and administrative, while the 12 post-treatment attributes are categorized as surgery, chemotherapy, and radiotherapy. Details on information extraction pertinent to natural language processing (NLP) can be seen in our work on smart extraction and analysis system for clinical research [29]. However, in the following sections, descriptions of the attributes that contribute to knowledge building are provided. Prior to storing information in the IDB, it is assessed for quality and standardization.

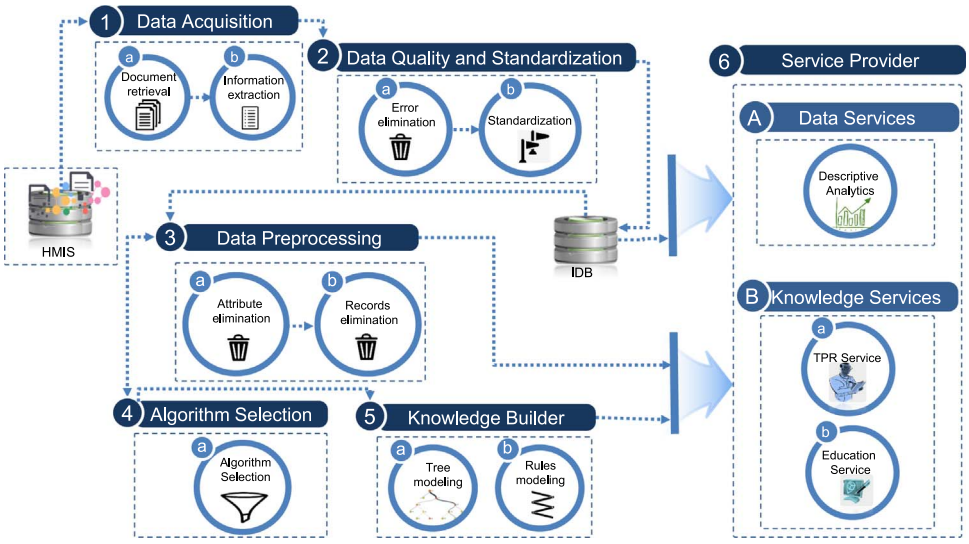

Fig. 1. Functional workflow diagram of the proposed methodology. HMIS=Health management information system, TPR=Treatment plan recommendation, IDB=Intermediate database.

3.2. Data quality and standardization

3.2.1. Data quality assessment (error elimination)

We designed staging value correction algorithm (Algorithm 1) that automatically perform corrections in the data. We used TNM classification (T: Primary tumor, N: Regional lymph nodes, and M: Distant metastasis) guidelines of the American Joint Committee on Cancer (AJCC) [30] to help in the corretion of clinical stage data. The proposed algorithm identifies and corrects the record if.

- one of the four values is left empty,
- one of the four values is entered incorrectly and there is no fuzziness in the record.

Fuzziness means if more than one values are candidates for the correctness. For instance, if we have an incorrect record in the data as shown below.

T	N	M	S	Status
T1	N0	M0	III	Incorrect

Mapping with guidelines, this record is identified as incorrect but there exist fuzziness in deciding which value is incorrect status as there are three possibilities of incorrectness: either S value is incorrect or T value, or N value. If either of them is corrected the whole record will become corrected as shown below. The encircled values show the correct values.

T	N	M	S	Status
T1	N0	M0	(I)	Correct
(T3)	N0	M0	III	Correct
T1	(N1)	M0	III	Correct

We did not automate the correction of such fuzzy records as without knowledge of clinical importance, it is hard to say which one to choose. One of the possible ways, the system shall direct such records to the concern user to verify by tracing from the source of first entries of the corresponding department(s).

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]

Table 2
Specification of attributes retrieved from a hospital management information system (HMIS) through the data acquisition function.

Pre-treatment attributes		
Demographic attributes	**Risk factors attributes**	**Diagnosis attributes**
1. Sex	1. Smoking	1. Site
2. Date of birth	2. Pan (a type of tobacco that is chewed)	2. Subsite
3. Ethnicity	3. Naswar (a moist, powdered tobacco snuff)	3. Histology
	4. Alcohol	4. Grade
Clinical Staging attributes	**Treatment attributes**	**Administrative attributes**
1. Tumor stage	1. Treatment intent	1. Hospital number
2. Node stage	2. Treatment plan	2. Date of clinic
3. Metastasis stage		3. Date of biopsy
4. Final clinical stage		4. Patient status
		5. Treatment status
Post-treatment attributes		
Surgery	**Chemotherapy**	**Radiotherapy**
1. Surgery intent	1. Regimen	1. Dose
2. Surgery procedure 1	2. No. of cycles	2. Fractions
3. Surgery procedure 2	3. Response	3. Response
4. Surgical margins		4. PEG (percutaneous endoscopic gastrostomy)
5. Lymph nodes		

Algorithm 1. TNM staging values correction.

$[T, N, M, S, TNMStageDBRecord]$
$T == \{Tis, T1, T2, T3, T4a, T4b, TAny\}$
$N == \{N0, N1, N2, T3, NAny\}$
$M == \{M0, M1\}$
$S == \{0, I, II, III, IVA, IVB, IVC\}$
$TNMStageDBRecord == set\ TNM\ Staging\ existing\ records\ of\ patients\ in\ the\ target\ database$
$TNM == T \times N \times M;\ SNM == S \times N \times M$

$TNMStageS : TNM \nrightarrow S;\ TNMStageT : SNM \nrightarrow T$
$TRecord : \mathbb{F}\ T$
$TValueOccurance : T \rightarrow N$

$SValue = ranTNMStageS$
$TValue = ranTNMStageT$
$TNMValues = domTNMStageS$
$SNMValues = domTNMStageT$
$TDBValues = ranSNMDBRecFunc$
$SNMDBRecord \cup ranTDBValues \subseteq TNMStageDBRecord$
$SNMDBRecord = domSNMDBRecFunc$
$TRecord = domTValueOccurance$

FindStageValue

$tnm? : TNM$
$stage! : S$

$stage! = tnm \in TNMValues \bullet TNMStageS(tnm?)$

FindTValue

$snm? : SNM$
$t! : T$

$t! = \{snm? \in SNMValues \bullet TNMStageT(snm?)\}$
$TDBValues = \langle TDBValues \upharpoonright \{n : N \bullet (n, t!)\}\rangle$
$TValueOccurance = TValueOccurance \cup \forall_v : T \mid v \in ranTDBValues \bullet \{v \mapsto \#(TDBValues \upharpoonright \{n : N \bullet (n, v)\})\}$
$t! = \exists_v 1 : T \mid v1 \in TRecord \bullet \forall_v 2 : T \mid v2 \in TRecord \wedge v1 \neq v2 \bullet TValueOccurance(v1) \geq TvalueOccurance(v2)$

Description of the algorithm: In Algorithm 1 which is formally represented using set theory and first-order predicate logic, the incorrect values are identified using the "FindStageValue" and "FindTValue" functions. FindStageValue identifies the missing/incorrect value of S and it is comparatively straightforward as it always yields single value for any patterns of TNM. So for as calculating T, N, or M is a bit different. Taking T as an example, for any given pattern of SNM, there is possibility that TNM staging guideline have more than one T candidate values. In this particular case, "FindTValue" formal function work as follows;

- All possible T values (t!) are identified for given SNM patterns from TNM guidelines.
- Retrieve the T values (TDBValues) for all patients which match the set of T values (t!) and SNM.
- Calculate the frequency of occurrences of each value in (t!).
- Choose the T value from (t!) as final value having maximum occurrences in the TDBValues.Note that "FindTValue" function can be replicated for finding N value and find M value because they have similar nature. In other words, "FindTValue" can be replaced with "FindNValue" to identify N value for the input of TMS values and "FindMValue" to identify M value for the input of TNS values. Other changes like "t!" shall be replaced with "n!" and "m!" accordingly.

3.2.2. Standardization of the language

There are two main types of variability are found in the data: domain variants and name variants, which were required to be represented in a standard language. Domain variants include mentions used by clinicians in clinical documents during practice. It is not necessary for variants to be recognized globally; rather, they are usually used in the EHR as localized terms. For example, the domain concept "radiotherapy" has domain variants of "radiation," "radiation therapy," "RT," and "RTx." Similarly, for name variants that are commonly used in clinical practice, we identified concepts that can be written under different names based on variations in spaces, hyphens (-), or numerals. For instance, for the clinical staging value "Clinical Stage 1," a variant "Clinical Stage I," is generated according to the heuristic of converting Arabic numerals to Roman numerals. The same can appear without a space "Clinical StageI" or with a hyphen "Clinical Stage-I."

To deal with the standardized language issue, we developed a dictionary-matching approach called the language standardization algorithm (LSA), as formalized in Algorithm 2. It is important to note that dictionary can be a locally developed domain dictionary or a standard dictionary derived from global domain such as SNOMED CT or UMLS. For this implementation, we used both local domain dictionary and a dictionary derived from SNOMED CT for a subset of head and neck cancer. LSA is a partial function of the DomainNameConcept, which is composed of StandardConcept and VariantConcept. The StandardConcpet always depends on the range of LSA, while the VariantConcept depends on the domain of LSA. The FindStandardConcept operation matches the InputVariant in the VariantConcept; if a match is found, the corresponding StandardConcept is returned from the LSA function; otherwise, the same InputVariant is returned.

Algorithm 2. Language standardization algorithm.

$[DomainNameVaiant]$
$LSA : DomainNameVaiant \nrightarrow DomainNameVaiant$
$StandardConcept : \mathbb{F}\ DomainNameVaiant$
$VariantConcept : \mathbb{F}\ DomainNameVaiant$

$StandardConcept = ranLSA$
$VariantConcept = domLSA$
$StandardConcept \cap VariantConcept = \emptyset$

FindStandardConcept

$InputVariant? : DomainNameVaiant$
$OutputStandardConcept! : DomainNameVaiant$

$OutputStandardConcept! = LSA(InputVariant?)$
$OutputStandardConcept! = \emptyset \Rightarrow OutputStandardConcept! = InputConcept?$

The corrected and standardized data are stored in the appropriate columns of IDB tables as structured data. Data services that mainly include descriptive analytics utilize the structured data directly from the IDB. The knowledge services that include recommendation services and education services can be applied to preprocessed data. More details on information extraction pertinent to natural language processing (NLP) can be seen in our work on smart extraction and analysis system for clinical research [29].

3.3. Data preprocessing

For knowledge services built on the data-driven approaches, the data need to be preprocessed before they are input in the machine learning algorithm. One of the preprocessing steps in data-driven approaches is dimensionality reduction [12], which mainly includes two types of reductions: attribute elimination and record elimination.

3.3.1. Attribute elimination

We employ three steps for attribute elimination. Firstly, we eliminate all of the administrative attributes such as hospital number, date of clinic visit, patient status, and treatment status. Secondly, we eliminate the irrelevant attributes that do not contribute to the predictive modeling; these mainly include the demographic attributes such as ethnicity and post-treatment attributes such as surgery, chemotherapy, and radiotherapy. Thirdly, we filter out all attributes that are missing more than 20% of their values, because they will most likely produce misleading results [31].

3.3.2. Record elimination

During data acquisition, we correct the TNM stage values for the records that have at most one incorrect value in the record. For example, a patient record has T=0, N=2, M=0, and S=5. In this record, the value for T and S are incorrect according to the AJCC TNM staging guidelines. All of the records with more than one incorrect/missing TNM stage value are removed from the dataset to avoid the chance of incorrect classification.

3.3.3. Missing data imputation

A missing data rate of less than 1% is trivial, 1–5% is manageable, 5–15% requires sophisticated methods to manage, and over 15% may seriously impact the overall interpretation [19]. The rate of missing data in our selected dataset fell within 1–5%, and we managed it by employing two

imputation techniques: guideline-based imputation and majority vote imputation [32]. Guideline-based imputation is applied to the staging attribute values, including tumor, node, metastasis, and final clinical stage values, while the majority vote is applied to the rest of the missing attribute values such as grade and subsite. We reused the FindStageValue and FindTValue functions of Algorithm 1 for the TNM stage missing value imputations. In the majority vote, first we identify the class of the instance that has a missing value. We collect all of the instances of the identified class and calculate the frequencies of each value. The value with the maximum frequency is substituted for the missing value. Overall 121 data values were found missing for different patient records. In Table 3, the description of imputation approach is provided for different attributes.

3.3.4. Preprocessed data specifications

After preprocessing, we are left with 13 attributes (12 independent and 1 dependent), as described in Table 4. The description column lists the meaning of each attribute and provides a list of values in parentheses. Fundamentally, the dependent attribute, i.e. treatment plan, has 10 classes; however, the classes (C: induction chemotherapy, S: surgery) with fewer records are excluded to avoid an imbalanced formation of the dataset. Four classes (RT, CRT, S RT, and C S CRT) are merged with four other classes (C RT, C CRT, C S RT, and C S CRT) based on the fact that all of the patients are given C treatment prior to any treatment but it is not mentioned explicitly during the entries. Even after merging, C S CRT fails to obtain the balance threshold value, so we exclude it from the dataset. Finally, three classes, i.e., C RT, C CRT, and C S RT, are included in the value set of the treatment plan attribute.At this stage, the proposed system facilitates the most frequent cases and does not cover the rare instances. However, this study can be extended to consider the rare classes which may involve oversampling mechanism or others similar techniques to create balance in the data.

3.4. Algorithm selection

As previously mentioned in this study, our target objective is to provide two knowledge-based services as applications:

- clinical decision support
- clinical education supportTo achieve this objective, we investigate an appropriate machine learning method. In the literature, different methods are reported to exhibit different characteristics, and they are mostly compared on the basis of model performance and comprehensibility [17]. Performance is a quantitative measure that can be deduced from measurements such as accuracy, number of involved attributes, and computational cost. Comprehensibility, on the other hand, is a subjective measure that is assessed by participating domain experts [17]. Aligned with our application objectives, we choose from the comprehensibility criteria such as model representation to disclose the inner workings and explanation abilities of the decisions. Another important aspect is the model output (rules) measurement. Clinicians would ideally choose a rule that demonstrates content validity [33]. In other words, the items in the rule are to be clinically sensible, which means that no obvious items are missing. Essentially, if a model generates too few rules, the output will be over-generalized. Similarly, too many rules will result in over-specialization. Over-generalization may hide the internal details of the model, while over-specialization may result in overfitting, which creates complications in the maintenance of the rules.

Table 3
Details of missing data imputations.

Attribute	No. of missing values	Imputation method
Grade	13	Resolved with majority vote method (mode)
Tumor (T)	69	Identified with guideline-based approach and resolved with majority vote (mode)
Node (N)	17	Identified with guideline-based approach and resolved with majority vote (mode)
Metastasis (M)	4	Identified with guideline-based approach and resolved with majority vote (mode)
Final Stage (S)	18	identified and resolved with majority vote (mode)

Table 4
Description of the attributes in the dataset used for prediction model development.

Attribute	Description
Sex	Indicates gender (male, female)
Grade	Indicates patient status (well, poor, moderate)
Treatment Intent	Patient,status for treatment (palliative or radical)
Clinical Stage T	TNM Staging T value (Tis,T1,T2,T3,T4a, T4b,TAny ()
Clinical Stage N	TNM Staging N value (N0,N1,N2,T3, NAny)
Clinical Stage M	TNM Staging M value (M0,M1)
Clinical Stage S	TNM Final Staging S value (0,I,II,III,IVA, IVB,IVC)
Smoking	Smoking status (yes, no)
Alcohol	Alcohol status (yes, no)
Naswar	Naswar status (yes, no); naswar is a moist, powdered tobacco snuff.
Pan	Pan status (yes, no); pan is type of tobacco that is chewed and finally spat out or swallowed
Histology	Indicate patient disease such as Squamous cell carcinoma, Adenoid cystic carcinoma, etc.
Treatment plan (Dependent attribute)	Treatment plan for patient (C, S, RT, CRT, C RT, C CRT, C S RT, S RT, S CRT, C S CRT), Where; C: induction chemotherapy, S: surgery, RT: radiotherapy, CRT: concurrent chemoradiation, S RT: S followed by RT, S CRT: S followed by CRT, C RT: C followed by RT, C CRT: C followed by CRT, C S RT: C followed by S RT, and C S CRT: C followed S CRT.

In the clinical domain, particularly in clinical decision support and education support applications, where the domain experts are expected to examine the internal workings of the model and determine how the decision is being made, we can choose techniques that are easy to understand and white box. Decision trees are powerful white-box classification techniques [34,35] with respect to expression capabilities, even though other white-box classification models such as k-nearest neighbors and logistic regression are also available. Decision trees not only provide explanation capabilities, but are also helpful for generating rules to be integrated with rule-based recommendation systems.

To achieve this desirable factor, the most commonly used algorithms are considered in the categories of decision trees (CRT/CART (Classification And Regression Trees) [36], CHAID (Chi-squared Automatic Interaction Detection) [37], J48/C4.5 [38], Quest [39], and LADTree (Logical Analysis of Data (LAD) Tree) [40,41] and decision rules (PART [42], Decision Tables [43], Ridor (Ripple DOwn Rules) [44,45], and JRip (RIPPER) [46]).

For evaluation of the algorithms, we first translate the qualitative criteria to quantitative criteria for ranking purposes. We use the weighted sum model (WSM) as expressed in Equation (1) for the selected criteria, where P is the accuracy of the algorithm, R is the number of rules, A is the number of attributes involved in the model, and U is the understandability of a model.

Equation 1: Weighted Sum Model (WSM) Score

$$A_i^{WSM-Score} = \alpha \sum_{j=1}^{m} w_j a_{ij}, \quad \text{for } i = 1, 2, 3, \ldots, m$$

where;

$$\begin{Bmatrix} \text{Criterian}(C_j) & \text{Accuracy(P)} & \text{Scaled} - \text{Rules}(R') & \text{Attributes(A)} & \text{Understandability(U)} \\ w_j & 0.5 & 0.1 & 0.1 & 0.3 \\ \beta_{ij} & 1 & 5 & 7.69 & 50 \end{Bmatrix}$$

\Here, α: is 0.01 and is a scaling constant to keep the final score value in the range between 0 and 1; w_j denotes the relative weight of the importance of the criterion C_j; a_{ij} is the performance value of algorithm A_i when it is evaluated in terms of the criterion; β_{ij} is the scaling constant of the

performance used in a_{ij} to keep the values in the range of 0 and 100; and $A_i^{WSM-Score}$ is the overall score of an algorithm, A_i, in terms of importance.

The values for w_j are assigned to different criteria based on importance. The performances, a_{ij} for P and A are from our experiment performed in the Weka environment [47]. Details of Weka experimentation are given in Table 5. The performance, a_{ij}, of R is scaled to R' as described in Eq. (2), and the performance, a_{ij}, of U is calculated as described in Eq. (3). The values of U are assigned based on the majority rule [48] involving four participants (domain experts). The participants were given the option of assigning U values as: 2, if the algorithm provides a high level of understandability, and 1 otherwise.

Let $[R_{min}, R_{max}]$ denote the acceptable number of rules where $R_{min} = 10$ and $R_{max} = 30$ are chosen based on the domain expert recommendation, R_A is the number of rules in Algorithm A, and the Scaled-Rule R' values are calculated as follow:

Equation 2: Scaled-Rule R' values assignments

$$R' = \begin{cases} (R_{max} - R_{min}) & \text{if } R_A \in [R_{min}, R_{max}]; \\ (2R_{max} - R_{min} - R_A) & \text{if } R_A > R_{max}; \\ 2R_A & \text{if } R_A < R_{min}. \end{cases}$$

Similarly, the understandability, U, is obtained as:

Equation 3: Understandability values assignments.

$$U = \begin{cases} 2 & \text{if } algorithm\ A \in Decision\ Tress; \\ 1 & \text{if } algorithm\ A \in Decision\ Rules. \end{cases}$$

We obtained values for the performances of each criterion as shown in Table 6 and found CRT decision tree algorithm to be the winner based on its highest rank value of 0.830. The Quest algorithms showed to second best performance with a slightly lower accuracy (0.2) than CRT.

3.5. Knowledge building

The CRT algorithm is selected to build knowledge for knowledge-based services: clinical decision support and clinical education support. The input data for building the knowledge model was the preprocessed structured data described in Table 4. In Fig. 2, we present the CRT generated model in the form of a visual tree diagram. For better visibility, the tree diagram is generated using an SPSS [49] tool without affecting the parameter setting of the Weka experiment. The tree shows the complete model consisting of 11 decision paths each starting from a parent node (node 0) and ending on terminal nodes (3, 7, 8, 10, 11, 15, 16, 17, 18, 19, and 20). A complete summary of the model is represented in Fig. 3.

As described in Fig. 3, the model includes 11 terminal nodes that represent the decision nodes. In other words, a total of 11 decision paths can be derived from the tree, and each decision path is then transformed into a rule. Finally, a total of 11 rules are contributing to building the knowledge base to be used in the clinical decision support service. It is pertinent to mention that the rules extracted with a machine learning approach might not be trusted by physicians, and the knowledge needs to be validated prior to use in a real system. One of the possible mechanisms for validation is described in our previous work [23] where the knowledge is validated with clinical guidelines provided by the National Comprehensive Cancer Network (NCCN) [50] for HNC. The validated knowledge is utilized to provide services of clinical decision support and education. It can be argued that, if extracted rules are required to be validated with NCCN, rules should be created directly from the NCCN guidelines. This is discussed in [23]; however, directly encoding from NCCN into rule-based form is not practicle due to its generic nature. Rules found out with statistical approaches are more granular than abstract guidelines. However, validation with guidelines will be necessary to adjust incorrect paths present in the extracted rules.

Table 5
Description of Weka experiment environment.

Parameter setting of algorithms in Weka environment					
Testing method: Cross-validation fold value = 10					
Decision tree algorithms			**Decision rules algorithms**		
Algorithm	**Option**	**Value**	**Algorithm**	**Option**	**Value**
J48	confidenceFactor	0.25	PART	confidenceFactor	0.25
	minNumObj	2		minNumObj	2
	numFolds	3		numFolds	3
	seed	1		seed	1
	subtreeRaising	True			
SimpleCART	heuristics	True	Decision Tables	crossVal	1
	minNumObj	2		evaluationMeasure	accuracy
	numFoldsPruning	5		search	BestFirst
	seed	1			
	usePrune	True			
LADTree	numOfBoostingIterations	10	Ridor	folds	3
				minNo	1
				shuffle	1

Table 6
Algorithm ranking based on criteria P, R', A, and U.

Classification model	Algorithm	P	$R \mapsto R'$	A	U	Ranking
Decision tree	CHAID	68.9	11 ↦ 20	6	2	0.791
	J48	66.45	43 ↦ 7	10	2	0.744
	Quest	68.8	10 ↦ 20	11	2	0.829
	CRT	69.0	11 ↦ 20	11	2	0.830
	LADTree	67.51	15 ↦ 20	6	2	0.784
Decision rules	PART	63.07	78 ↦ − 28	13	1	0.425
	Decision Tables	66.66	21 ↦ 20	4	1	0.614
	Ridor	65.18	34 ↦ 16	3	1	0.633

3.6. Service provider

Prior to mention the services designed over the structured data, it is pertinent to describe some relevant problems and experiences we faced during this study while designing and developing the data and knowledge services. The main issues are faced in the areas of data structure design and data preparation. Usually, the data analysis requires data distributed across multiple clinical documents which need to be logically integrated into a unified design in order to perform analysis comprehensively. Computer experts alone may not be able to complete this task without a close support of clinicians and researchers -the real stakeholders of the service. Secondly, data requirements for different services are different. For instance, machine learning based knowledge model needs data completion by filling missing value imputations which may not be required by other analysis approaches. Keeping that in view, in this study, we keep the original corrected data unaffected from the imputed data prepared for the knowledge model.

3.6.1. Data services

We developed a transformation function that transforms the data based on the user-provided query in an SPSS [49] compliant format. In SPSS, the trained users apply various kinds of filters to generate data for analysis. Some of the findings from these data are published in [51–53] by clinicians who are collaborators in this research. The interfaces of proposed head neck cancer (HNC) system with SPSS are described in Fig. 4. SPSS is interfaced with proposed system through two ways.

- ODBC Connection, SPSS Query Editor establishes ODBC connection directly with SQL server-based structured database.
- SPSS.NET Plugin, SPSS Transformation service of HNC system formats the SQL data in SPSS format and export to SPSS data repository.

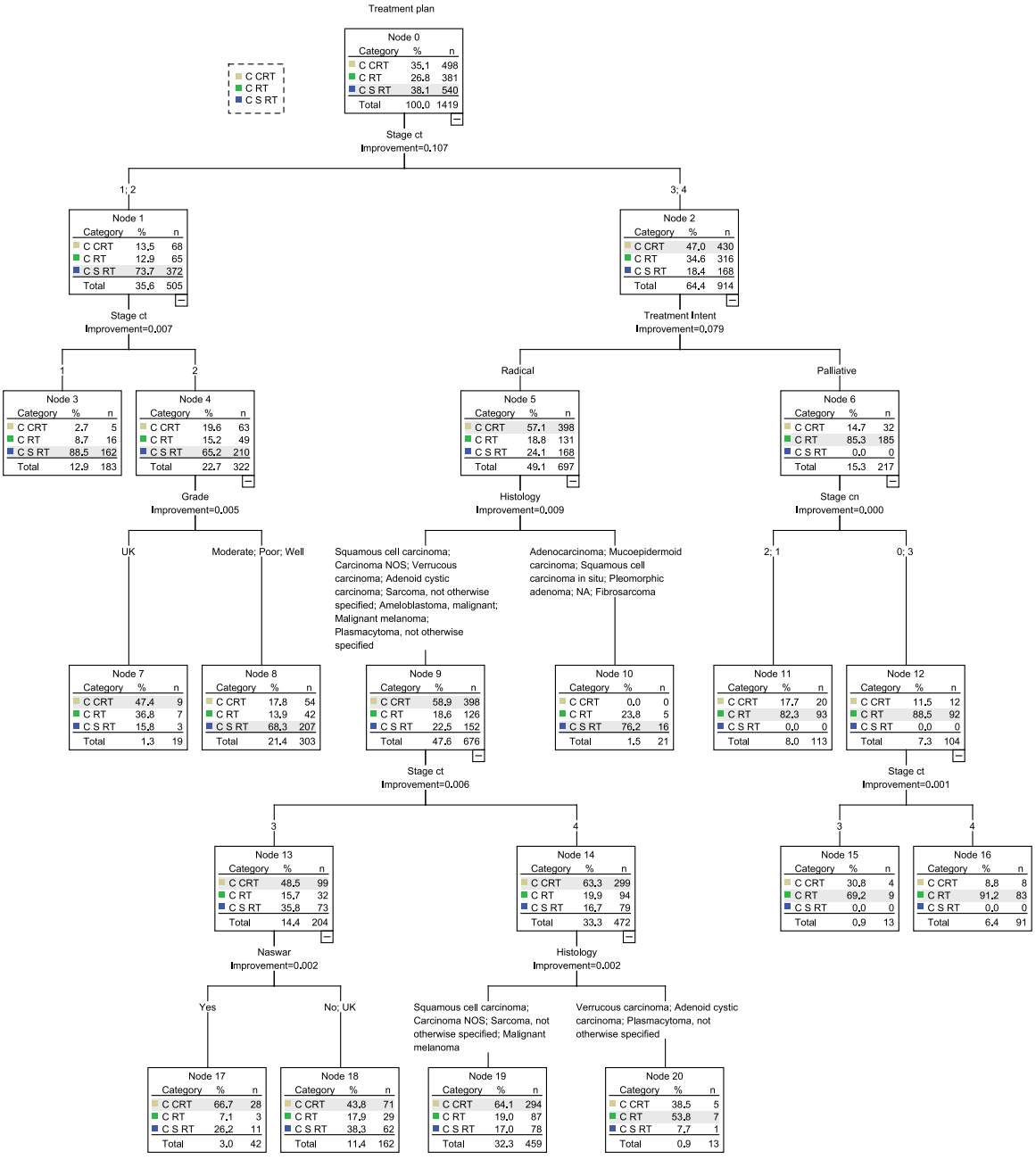

Fig. 2. CRT-generated decision tree model consisting of 21 total nodes 11 terminal nodes.

3.6.2. Knowledge services

The proposed methodology supports two types of knowledge services: clinical decision support (CDS) services and clinical education support (CES) services. The flat if-then rules are of great help to be transformed into a knowledge base (KB) of CDS. Knowledge in the form of a decision tree can potentially contribute to the development of CES services. Clinicians can examine the process of how a decision is being made for the given conditions of a patient by exploring the executed decision path in the decision tree. A decision path provides enough information to understand the important variables in the decision-making process. The knowledge services are interfaced with SPSS by integrating the CRT-generated decision tree model in the HNC system.

4. Implementation and results

4.1. Experimental environment

The system developed based on proposed methodology was deployed in a real environment, as a part of an in-house developed HMIS in Pakistan [54]. We performed two types of experiments: online and offline in order to acquire patient data from the source system to the target system. In online experiment, all the pertinent clinical documents are retrieved through the use of web service and required patient data is extracted. The acquired patient data is passed through quality and standardization functions prior to store in the structured database of the target system. While, in offline experiment, based on the system scope and objective, a subset of overall data is considered and preprocessed in order to prepare the data for the knowledge model creation. The description of the experimental environmental is provided in Fig. 5.

4.2. Dataset description

Based on the online experiment functions of data acquisition, we acquired a dataset consists of 3447 patient records with reference to 13 788 clinical documents. The descriptions of overall dataset that is

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	Treatment plan
	Independent Variables	Smoking, Pan, Naswar, Alcohol, Current event, Treatment Intent, Histology, Grade, Stage ct, Stage cn, Stage cm, Stage cs, Sex
	Validation	Cross Validation
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	10
Results	Independent Variables Included	Stage ct, Stage cs, Histology, Alcohol, Pan, Naswar, Smoking, Grade, Treatment Intent, Stage cn, Stage cm
	Number of Nodes	21
	Number of Terminal Nodes	11
	Depth	5

Fig. 3. Model summary with specifications and output results.

used in the system implementation and a subset of overall dataset that is used for knowledge model creation are provided in Table 7.

4.3. Error corrections

We present here the results of the TNM staging value corrections based on the Algorithm 1 of the data quality task. In Table 8, part (a) shows the number of records, in which we found 39 records with an error related to the final stage S value, (b) shows the 53 records with an incorrect T value, (c) shows the 38 records with an error in the N value, and (d) shows the M value errors in a total of 30 records.

4.4. Model accuracy

The accuracy of each class of dependent attribute "treatment plan" is presented in Table 9. The classes include C CRT, C RT, and C S RT. The C CRT class showed the highest accuracy of 80.7%, followed by C S RT with 71.3% accuracy. The C RT class showed comparatively lower accuracy, possibly due to the lower prior probability (0.268) obtained from the training sample compared to the prior probabilities of C CRT (0.351) and C S RT (0.381).

The decision paths in a tree model (Fig. 2) consists of five levels with 11 terminal nodes and 21 total nodes. At the first level, the tree is divided into two groups on the basis of Tumor T values of 2:1 and 4:3, where 505 patients belonging to the first group (2:1) are given the C S RT treatment plan with 73.7% accuracy and 914 patients belonging to the second group (4:3) are given the C CRT treatment plan with 47.0% accuracy. Further division of group 4:3 is made on the basis of treatment intent, where 697 patients having a radical intent are provided C CRT with 57.1% accuracy, and the 217 patients having a palliative intent are provided the C RT treatment plan with 85.3% accuracy. In the same manner, the tree has grown to the third level, and divisions are made on the basis of histology, grade, and Node N stage. At the fourth level, divisions are made based on Tumor T; finally, the division is made on the basis of the risk factors Naswar (a moist, powdered tobacco snuff) and histology.

Eleven overall decision paths are drawn from the entire classification tree consisting of nodes and terminal nodes. Table 10 lists the decision paths with the node description, predicted treatment plan, number of patients, and corresponding accuracy.

The corresponding rules derived from the decision paths in Table 10 are provided in Table 11. For brevity, we used histologies A, B, C, and D, which in reality are represented as groups of values, as follows:

A = Squamous cell carcinoma; carcinoma NOS; Verrucous carcinoma; adenoid cystic carcinoma; Sarcoma, not otherwise specified; Ameloblastoma, malignant; Malignant melanoma; Plasmacytoma, not otherwise specified.

B = Adenocarcinoma; Mucoepidermoid carcinoma; Squamous cell carcinoma in situ; Pleomorphic adenoma; NA; Fibrosarcoma.

C = Squamous cell carcinoma; carcinoma NOS; Sarcoma, not otherwise specified; Malignant melanoma.

D = Verrucous carcinoma; adenoid cystic carcinoma; Plasmacytoma, not otherwise specified.

At a more granular level, a patient is given a treatment plan on the basis of conditions explained as follows:

1. The C S RT treatment plan is recommended for patients with a tumor T stage value of 1 or 2; grade values of poor, moderate, or well; and a histology value of adenocarcinoma, mucoepidermoid carcinoma, squamous cell carcinoma in situ, pleomorphic adenoma, NA, or fibrosarcoma.
2. The C CRT treatment plan is suggested for radical patients only if they have a clinical stage T value of 2, 3 or 4; one of the following histology values: squamous cell carcinoma, carcinoma NOS, adenoid cystic carcinoma, verrucous carcinoma, malignant melanoma, sarcoma, or plasmacytoma, grade value UK (unknown), and a Naswar risk factor value of 'yes'.
3. The C RT treatment plan is recommended for palliative patients only if they have a clinical stage T value of 3 or 4; a histology value of one of the following: squamous cell carcinoma, carcinoma NOS, adenoid cystic carcinoma, verrucous carcinoma, malignant melanoma, sarcoma, or plasmacytoma; a treatment intent value of either radical or palliative; and a Node N stage value of either 0, 1, 2, or 3.

5. Discussion

5.1. Implication and applications

The techniques described in this study can potentially contribute to the development of clinical research and CDS systems if properly affiliated with an organization's objectives, because the success of a CDSS greatly depends on its capability to be integrated into a health information system (HIS) [54]. Based on the experience gained through the experiments of system implementation and integration in a hospital environment, we observed that the data acquisition from clinical documents provides a great opportunity to create a prediction model to assist in decision making and clinical research. The prediction model helps not only in the decision-making process for expert physicians, but also in the education of inexperienced

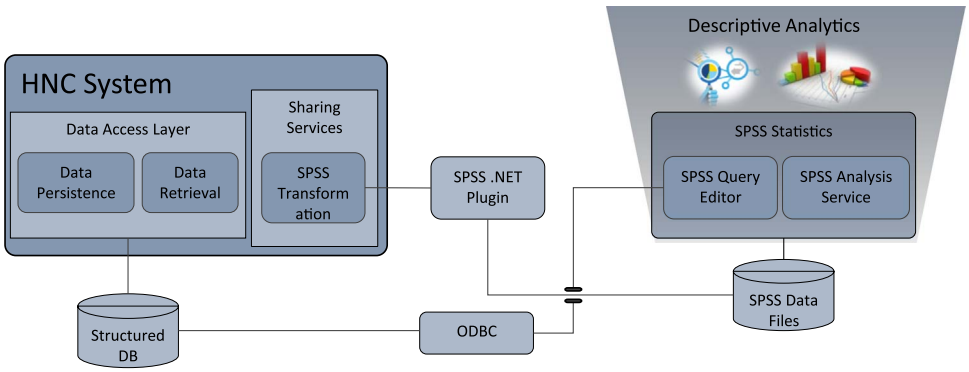

Fig. 4. Head neck cancer (HNC) system interfaces with SPSS analysis tool.

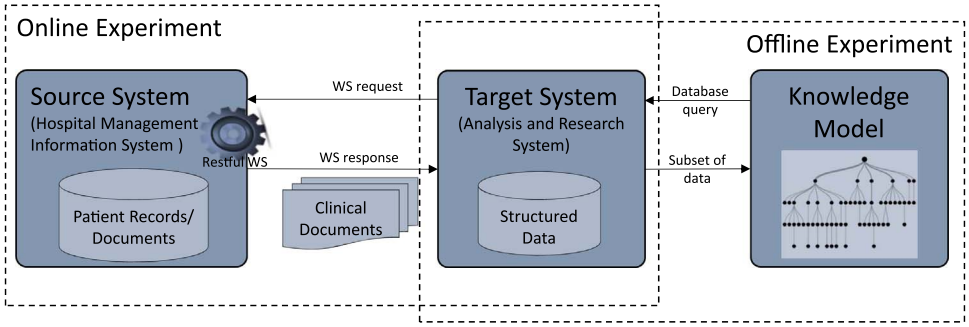

Fig. 5. Representation of experimental environment for online and offline experiment.

clinicians including resident doctors and nurses. Such inexperienced clinicians can learn how a decision is made for the given conditions of patients by exploring the executed decision path. A decision path provides enough information to understand the important variables in the decision-making process.

5.2. Generalization of the methods

The prediction model discussed in Section 3.5 was tested on data from patients with cancer of the oral cavity. However, the retrieved dataset also includes data for patients with cancer in other sites such as the salivary glands, nasopharynx, and larynx. A number of useful statistical results can be derived from the structured data. Some of the findings from these data are published in [52,51,53] by physicians who are collaborators in this research. Similarly, the algorithms developed for the TNM staging correction and language standarization are reusable for other sites without any modification to the structure except for updating the TNM guidelines tables and standard dictionaries. Moreover, integrating the proposed system with EHR, the RESTful services and JSON implementation is not required to be changed technically except the URLs and URIs for the service methods developed at the EHR side. On the other hand, the schema designed for holding the structured data is flexible enough to be reused unless the data scope remain unchanged. Data scope means the inclusion of new types of data from the EHR side.

5.3. Limitations of the work

There are some limitations to this work. Firstly, we only looked at clinical documents from one hospital, and we were unable to compare the results with existing systems due to the extensive amount of customization required to implement the existing systems for a custom domain, i.e., HNC. Secondly, we tested the prediction model for only one site, “oral cavity” of HNC, even though we extracted data from other anatomical sites such as the salivary glands and pharynx. This choice was made because we had access to support from domain experts specialized in the oral cavity site. It is thus important to assess the performance of our system with a document set selected from alternative source and extending the prediction model to the data of other sites. More specifically, the customization will be required in the following areas.

Table 7
Dataset description.

No. of total patient records: 3447 - No. of oral cavity records: 1526 - No. of eliminated records: 106 - No. of records used for model creation: 1420	No. of total documents: 13788 - Structured: 3447 - Semi-structured: 9872 - Structured: 469 - No. of documents used for model creation: 3052
No. of total attributes: 34 - Pre-treatment: 22 - Post-treatment: 12 - No. of eliminated attributes: 20 - No. of attributes used for model creation: 14	Note: Individual details of attributes can be seen in Table 2 and Table 3.

- Data acquisition, technically RESTful services and JSON implementation is generic, however, queries run at the backend of service need to be customized according to the service requirements.
- Information extraction, based on the domain change, localized and standard dictionaries need to be updated in order to recognized the named entities of the changed domain.
- Data preprocessing, the functions such as missing data imputation need to checked for the data of changed domain. Depends on the data, the simple imputation techniques may be extended to the multiple imputation technique.

Moreover, the knowledge model is built using a single machine learning method which can be further enhanced to be trained using a combination of methods (ensemble learning) in order to get a better accuracy.

5.4. Lessons learned

Clinical decision support system require two essential ingredients in order to produce intended results. One of these two ingredients is data which has many factors to consider prior to develop any sort of knowledge model over it. We want to discuss few of those factors which

Table 8
No. of incorrect values of each S, T, N, and M for oral cavity site.

(a) No. of final stage S		(b) No. of tumor T	
Stage values	**Incorrect/Total records**	**Tumor values**	**Incorrect/Total records**
Stage 0	0/0	Tis	0/0
Stage I	6/62	T1	6/183
Stage II	6/225	T2	11/322
Stage III	8/260	T3	12/326
Stage IV A, B, C	19/772	T4 a, b, c	24/688
Total	39/1419	Total	53/1419

(c) No. of node N		(d) No. of metastasis M	
Node values	**Incorrect/Total records**	**Metastasis values**	**Incorrect/Total records**
N0	29/846	M0	30/1417
N1	7/269	M1	0/2
N2	1/260	Total	30/1419
N3	1/44		
Total	38/1419		

Table 9
By class accuracy of treatment plan prediction model built with CRT decision tree algorithm.

Observed	**Predicted**			
	C CRT	C RT	C S RT	Percent correct (%)
C CRT	402	37	59	80.7
C RT	126	192	63	50.4
C S RT	154	1	385	71.3
Overall percentage				69.0

Table 10
Decision paths derived from testing prediction model (TP=Treatment Plan).

Paths	Nodes included	TP (Dominant)	Patients	Path Accuracy (%)
Path 1	Node 0, Node 1 → Terminal Node 3	C S RT	162	88.5
Path 2	Node 0, Node 1, Node 4 → Terminal Node 7	C CRT	9	47.4
Path 3	Node 0, Node 1, Node 4 → Terminal Node 8	C S RT	207	68.3
Path 4	Node 0, Node 2, Node 5, Node 9, Node 13 → Terminal Node 17	C CRT	28	66.7
Path 5	Node 0, Node 2, Node 5, Node 9, Node 13 → Terminal Node 18	C CRT	71	43.8
Path 6	Node 0, Node 2, Node 5, Node 9, Node 14 → Terminal Node 19	C CRT	294	64.1
Path 7	Node 0, Node 2, Node 5, Node 9, Node 14 → Terminal Node 20	C RT	7	53.8
Path 8	Node 0, Node 2, Node 5 → Terminal Node 10	C S RT	16	76.2
Path 9	Node 0, Node 2, Node 6 → Terminal Node 11	C RT	93	82.3
Path 10	Node 0, Node 2, Node 6, Node 12 → Terminal Node 15	C RT	9	69.2
Path 11	Node 0, Node 2, Node 6, Node 12 → Terminal Node 16	C RT	83	91.2

Table 11
Decision rules derived from decision paths.

Rules	Conditions → Decision
Rule 1	Stage ct = (1 or 2) AND Stage ct = 1 → Treatment Plan=C S RT
Rule 2	Stage ct = (1 or 2) AND Stage ct = 2 AND Grade = UK → Treatment Plan = C CRT
Rule 3	Stage ct = (1 or 2) AND Stage ct = 2 AND Grade = (Moderate or Poor or Well) → Treatment Plan = C S RT
Rule 4	Stage ct = (3 or 4) AND Treatment Intent = Radical AND Histology = A AND Stage CT = 3 AND Naswar = Yes → Treatment Plan = C CRT
Rule 5	Stage ct = (3 or 4) AND Treatment Intent = Radical AND Histology = A AND Stage CT = 3 AND Naswar = (No or UK) → Treatment Plan = C CRT
Rule 6	Stage ct = (3 or 4) AND Treatment Intent = Radical AND Histology = A AND Stage CT = 4 AND Histology = C → Treatment Plan = C CRT
Rule 7	Stage ct = (3 or 4) AND Treatment Intent = Radical AND Histology = A AND Stage CT = 4 AND Histology = D → Treatment Plan = C RT
Rule 8	Stage ct = (3 or 4) AND Treatment Intent = Radical AND Histology = B → Treatment Plan = C S RT
Rule 9	Stage ct = (3 or 4) AND Treatment Intent = Palliative AND Stage cn = (1 or 2) → Treatment Plan = C RT
Rule 10	Stage ct = (3 or 4) AND Treatment Intent = Palliative AND Stage cn = (0 or 3) AND Stage ct = 3 → Treatment Plan = C RT
Rule 11	Stage ct = (3 or 4) AND Treatment Intent = Palliative AND Stage cn = (0 or 3) AND Stage ct = 4 → Treatment Plan = C RT

we have experienced during the course of analysis, design, and development of proposed system.

- Hospital management/information systems are not very flexible to generate decision and educational support services in a straight manner. There we need to study the inherent semantics of the system generated data which may be in structured, semi-structured, or completely in unstructured format in addition to the technical aspects of integration of newly research system with the source hospital information system. What type of documents, containing what kind of data and particularly what portion(s) of a document are relevant to be extracted? Are any sort of standard vocabulary is used and at what level: category level or individual value level? Are there any abbreviations used that may or may have any presence in global dictionaries such as, WordNet? These type of very higher level but key questions are needed to satisfy in the requirement elicitation and analysis phase. Finding answer to these questions reveal interesting implication to help in determining dependencies that may exist in the data which in terms help in error correction and language standardization.
- Data dependencies exist in the data are implicit most of the times which may provide semantically incorrect results at the end. For instance, TNM staging involves four attributes: tumor, node, metastasis, and clinical stage. Initial level of knowledge is acquired from the domain experts in order to know the basic level of understanding and link to the reference guidelines. Based on TNM staging reference guidelines, we developed data correction algorithm. Similarly, understanding of vocabulary and abbreviation led us to standardize the language of data used for research analysis and decision support.
- In order to create a better knowledge model, a very clear objective of the system need to be set up. In objective it is identified the system is aimed to support in diagnosis, prognosis, etiology, or treatment. Based on the objective, a subset of overall dataset is identified. At this level, more focus should be given to resolving the imbalance distribution of data, missing values in the data, and selection of appropriate machine learning (ML) method. It is pertinent to emphasize that ML method selection should not be based only on statistical measurements rather nature of the data, objective of the system, and requirements of the user in terms of explanation for why and how questions. In clinical domain, physicians majorly want explicit results which means they need to know why this decision is made and how it is made.

To answer these questions, strong tie between technical engineers and domain experts plus independent analysis is required. In this study, we have long-way collaboration with physicians of SKMCH & RC hospital and multiple meeting sessions we were able to get answer for these questions.

6. Conclusion

Electronic health records provide invaluable information for educating and enabling clinicians in informed decision making. EHRs are under-utilized due to the lack of automatic knowledge acquisition methods. Previous work in this domain mainly concentrated on the descriptive statistics of the health records. Our proposed CKM-CT methodology employs automatic data acquisition from clinical documents, data quality, language standardization, preprocessing, and machine learning algorithm selection. The proposed approach is realized for the domain of head and neck cancer, however, it can be applied to the data of other domains such as breast cancer with similar objectives. In the future, we plan to replicate the proposed methods for other sites of head and neck cancer, e.g., the salivary glands and pharynx.

Conflict of interest

None declared.

Acknowledgements

This work was supported by the Industrial Core Technology Development Program (10049079, Development of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea)" and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) NRF-2014R1A2A2A01003914. The research was also supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support

program (IITP-2015-(H8501-15-1015) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

[1] A. Jemal, R. Siegel, J. Xu, E. Ward, Cancer statistics, 2010, CA: Cancer J. Clin. 60 (5) (2010) 277–300.

[2] V. Slavov, P. Rao, S. Paturi, T.K. Swami, M. Barnes, D. Rao, R. Palvai, A new tool for sharing and querying of clinical documents modeled using hl7 version 3 standard, Comput. Methods Progr. Biomed. 112 (3) (2013) 529–552.

[3] I. Spasić, J. Livsey, J.A. Keane, G. Nenadić, Text mining of cancer-related information: review of current status and future directions, Int. J. Med. Inform. 83 (9) (2014) 605–623.

[4] J. Mathe, J. Sztipanovits, M. Levy, E.K. Jackson, W. Schulte, Cancer treatment planning: formal methods to the rescue, in: Proceedings of the 4th International Workshop on Software Engineering in Health Care, IEEE Press, Zurich, Switzerland, 2012, pp. 19–25

[5] M. Hewitt, J.V. Simone, et al., Ensuring Quality Cancer Care, National Academies Press, Washington, D.C., 1999.

[6] M. Karanikolos, L. Ellis, M.P. Coleman, M. McKee, Health systems performance and cancer outcomes, J. Natl. Cancer Inst. Monogr. 46 (1) (2013) 7–12.

[7] G. Mikkelsen, J. Aasly, Concordance of information in parallel electronic and paper based patient records, Int. J. Med. Inform. 63 (3) (2001) 123–131.

[8] G. Mikkelsen, J. Aasly, Narrative electronic patient records as source of discharge diagnoses, Comput. Methods Progr. Biomed. 71 (3) (2003) 261–268.

[9] F. Liu, C. Weng, H. Yu, Natural language processing, electronic health records, and clinical research, in: Clinical Research Informatics, Springer, London, 2012, pp. 293–310

[10] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (6) (2012) 395–405.

[11] N. Douali, H. Csaba, J. De Roo, E.I. Papageorgiou, M.-C. Jaulent, Diagnosis support system based on clinical guidelines: comparison between case-based fuzzy cognitive maps and bayesian networks, Comput. Methods Progr. Biomed. 113 (1) (2014) 133–143.

[12] L. Goodwin, M. VanDyne, S. Lin, S. Talbert, Data mining issues and opportunities for building nursing knowledge, J. Biomed. Inform. 36 (4) (2003) 379–388.

[13] D. Demner-Fushman, W.W. Chapman, C.J. McDonald, What can natural language processing do for clinical decision support?, J. Biomed. Inform. 42 (5) (2009) 760–772.

[14] E.S. Chen, G. Hripcsak, H. Xu, M. Markatou, C. Friedman, Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study, J. Am. Med. Inform. Assoc. 15 (1) (2008) 87–98.

[15] B. Pandey, R. Mishra, Knowledge and intelligent computing system in medicine, Comput. Biol. Med. 39 (3) (2009) 215–230.

[16] M. Torii, K. Wagholikar, H. Liu, Using machine learning for concept extraction on clinical documents from multiple data sources, J. Am. Med. Inform. Assoc. 18 (5) (2011) 580–587.

[17] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: current issues and guidelines, Int. J. Med. Inform. 77 (2) (2008) 81–97.

[18] A.W.-C. Liew, N.-F. Law, H. Yan, Missing value imputation for gene expression data: computational techniques to recover missing data from available information, Brief. Bioinform. 12 (5) (2011) 498–513.

[19] E. Acuna, C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, in: Classification, Clustering, and Data Mining Applications, Springer, Berlin Heidelberg, 2004, pp. 639–647.

[20] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Stat. Med. 30 (4) (2011) 377–399.

[21] K. Morik, P. Brockhausen, T. Joachims, Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1999.

[22] M. Skevofilakas, K. Nikita, P. Templaleksis, K. Birbas, I. Kaklamanos, G. Bonatsos, A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, Shanghai, China, 2005, pp. 2429–2432.

[23] M. Hussain, M. Afzal, T. Ali, R. Ali, W.A. Khan, A. Jamshed, S. Lee, B.H. Kang, K. Latif, Data-driven knowledge acquisition, validation, and transformation into hl7 arden syntax, Artif. Intell. Med., http://dx.doi.org/10.1016/j.artmed.2015.09.008

[24] J. Lopez, D. Palacios-Alonso, S. Tortajada, A. Moreno, E. Casitas, J. García-Gómez, R.G. Otal, A. Pérez-González, A. Martinez, C.P. Calderon, et al., Computerized decision support system and naïve bayes models for predicting the risk of relapse in breast cancer, Int. J. Radiat. Oncol. Biol. Phys. 90 (1) (2014) S593–S594.

[25] L. Moja, A. Passardi, M. Capobussi, R. Banzi, F. Ruggiero, K. Kwag, E.G. Liberati, M. Mangia, I. Kunnamo, M. Cinquini, et al., Implementing an evidence-based computerized decision support system linked to electronic health records to improve care for cancer patients: the onco-codes study protocol for a randomized controlled trial, Implement. Sci. 11 (1) (2016) 153.

[26] A. Stojadinovic, A. Bilchik, D. Smith, J.S. Eberhardt, E.B. Ward, A. Nissan, E.K. Johnson, M. Protic, G.E. Peoples, I. Avital, et al., Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model, . Surg. Oncol. 20 (1) (2013) 161–174.

[27] O. Kohany, A.J. Gentles, L. Hankus, J. Jurka, Annotation, submission and screening of repetitive elements in repbase: repbasesubmitter and censor, BMC Bioinform. 7 (1) (2006) 1.

[28] L. Richardson, S. Ruby, RESTful web services, " O'Reilly Media, Inc.", 2008.

[29] M. Afzal, M. Hussain, W.A. Khan, T. Ali, A. Jamshed, S. Lee, Smart extraction and analysis system for clinical research, Telemed. e-Health, http://dx.doi.org/10.1089/tmj.2016.0157

[30] S. Edge, D. Byrd, C. Compton, A. Fritz, F. Greene, A. Trotti, American Joint Committee on Cancer, AJCC Cancer Staging Manual 7.

[31] R. Ali, J. Hussain, M.H. Siddiqi, M. Hussain, S. Lee, H2rm: a hybrid rough set reasoning model for prediction and management of diabetes mellitus, Sensors 15 (7) (2015) 15921–15951.

[32] S.G. Liao, Y. Lin, D.D. Kang, D. Chandra, J. Bon, N. Kaminski, F.C. Sciurba, G.C. Tseng, Missing value imputation in high-dimensional phenomic data: imputable or not, and how?, BMC Bioinform. 15 (1) (2014) 1.

[33] I.G. Stiell, G.A. Wells, Methodologic standards for the development of clinical decision rules in emergency medicine, Ann. Emerg. Med. 33 (4) (1999) 437–447.

[34] S. Dreiseitl, L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inform. 35 (5) (2002) 352–359.

[35] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127.

[36] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, CRC Press, Boca Raton, Florida, 1984.

[37] Y.M. Chae, S.H. Ho, K.W. Cho, D.H. Lee, S.H. Ji, Data mining approach to policy analysis in a health insurance domain, Int. J. Med. Inform. 62 (2) (2001) 103–111.

[38] J.R. Quinlan, C4. 5: Programming for Machine Learning, Morgan Kauffmann, San Francisco, CA, USA, 1993

[39] W.-Y. Loh, Y.-S. Shih, Split selection methods for classification trees, Stat. Sin. (1997) 815–840.

[40] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, M. Hall, Multiclass alternating decision trees, in: Machine Learning: ECML 2002, Springer, Berlin Heidelberg, 2002, pp. 161–172.

[41] S. Thaseen, C.A. Kumar, An analysis of supervised tree based classifiers for intrusion detection system, in: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, Tamilnadu, India, 2013, pp. 294–299.

[42] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, in: ICML, vol. 98, 1998, pp. 144–151.

[43] R. Kohavi, The power of decision tables, in: Machine Learning: ECML-95, Springer, Berlin Heidelberg, 1995, pp. 174–189.

[44] P. Compton, G. Edwards, B. Kang, L. Lazarus, R. Malor, P. Preston, A. Srinivasan, Ripple down rules: turning knowledge acquisition into knowledge maintenance, Artif. Intell. Med. 4 (6) (1992) 463–475.

[45] P. Compton, G. Edwards, B. Kang, L. Lazarus, R. Malor, T. Menzies, P. Preston, A. Srinivasan, C. Sammut, Ripple down rules: possibilities and limitations, in: Proceedings of the Sixth AAAI Knowledge Acquisition for Knowledge-Based Systems Workshop, Calgary, Canada, University of Calgary, 1991, pp. 6–1.

[46] W.W. Cohen, Fast effective rule induction, in: Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 115–123.

[47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.

[48] J. Zhang, C.K. Hsee, Z. Xiao, The majority rule in individual decision making, Organ. Behav. Hum. Decis. Process. 99 (1) (2006) 102–111.

[49] I. SPSS, Decision trees 21 (2012).

[50] NCCN, National Comprehensive Cancer Network, http://www.nccn.org/, [Online; accessed 24-April-2015] (2015).

[51] H. Iqbal, A. Bhatti, R. Hussain, A. Jamshed, Ten year experience with surgery and radiation in the management of malignant major salivary gland tumors, Asian Pacific J. Cancer Prevent.: APJCP 15 (5) (2014) 2195.

[52] H. Iqbal, A.B.H. Bhatti, R. Hussain, A. Jamshed, Regional failures after selective neck dissection in previously untreated squamous cell carcinoma of oral cavity, Int. J. Surg. Oncol. (2014).

[53] A. Jamshed, R. Hussain, H. Iqbal, Gemcitabine and cisplatin followed by chemo-radiation for advanced nasopharyngeal carcinoma, Asian Pacific J. Cancer Prevent.: APJCP 15 (2) (2013) 899–904.

[54] F. Sultan, M.T. Aziz, I. Khokhar, H. Qadri, M. Abbas, A. Mukhtar, W. Manzoor, M.A. Yusuf, Development of an in-house hospital information system in a hospital in Pakistan, Int. J. Med. Inform. 83 (3) (2014) 180–188.