



Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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



Data-driven knowledge acquisition, validation, and transformation into HL7 Arden Syntax

Maqbool Hussain^a, Muhammad Afzal^a, Taqdir Ali^a, Rahman Ali^a, Wajahat Ali Khan^a, Arif Jamshed^b, Sungyoung Lee^{a,*}, Byeong Ho Kang^c, Khalid Latif^d

^a Department of Computer Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si 446-701, Gyeonggi-do, Republic of Korea

^b Department of Radiation Oncology, Shaikat Khanum Memorial Cancer Hospital and Research Centre, 7A Block R-3, M.A. Johar Town, Lahore 54782, Pakistan

^c Computing and Information Systems, University of Tasmania, Hobart 7001, Tasmania, Australia

^d Department of Computer Science, COMSATS Institute of Information Technology, Park Road, Islamabad 45550, Pakistan

ARTICLE INFO

Article history:

Received 1 January 2015

Received in revised form

15 September 2015

Accepted 15 September 2015

Keywords:

Knowledge acquisition

Knowledge validation

Prediction models

Clinical guidelines

Clinical decision support systems

HL7 Arden Syntax

ABSTRACT

Objective: The objective of this study is to help a team of physicians and knowledge engineers acquire clinical knowledge from existing practices datasets for treatment of head and neck cancer, to validate the knowledge against published guidelines, to create refined rules, and to incorporate these rules into clinical workflow for clinical decision support.

Methods and materials: A team of physicians (clinical domain experts) and knowledge engineers adapt an approach for modeling existing treatment practices into final executable clinical models. For initial work, the oral cavity is selected as the candidate target area for the creation of rules covering a treatment plan for cancer. The final executable model is presented in HL7 Arden Syntax, which helps the clinical knowledge be shared among organizations. We use a data-driven knowledge acquisition approach based on analysis of real patient datasets to generate a predictive model (PM). The PM is converted into a refined-clinical knowledge model (R-CKM), which follows a rigorous validation process. The validation process uses a clinical knowledge model (CKM), which provides the basis for defining underlying validation criteria. The R-CKM is converted into a set of medical logic modules (MLMs) and is evaluated using real patient data from a hospital information system.

Results: We selected the oral cavity as the intended site for derivation of all related clinical rules for possible associated treatment plans. A team of physicians analyzed the National Comprehensive Cancer Network (NCCN) guidelines for the oral cavity and created a common CKM. Among the decision tree algorithms, chi-squared automatic interaction detection (CHAID) was applied to a refined dataset of 1229 patients to generate the PM. The PM was tested on a disjoint dataset of 739 patients, which gives 59.0% accuracy. Using a rigorous validation process, the R-CKM was created from the PM as the final model, after conforming to the CKM. The R-CKM was converted into four candidate MLMs, and was used to evaluate real data from 739 patients, yielding efficient performance with 53.0% accuracy.

Conclusion: Data-driven knowledge acquisition and validation against published guidelines were used to help a team of physicians and knowledge engineers create executable clinical knowledge. The advantages of the R-CKM are twofold: it reflects real practices and conforms to standard guidelines, while providing optimal accuracy comparable to that of a PM. The proposed approach yields better insight into the steps of knowledge acquisition and enhances collaboration efforts of the team of physicians and knowledge engineers.

© 2015 Elsevier B.V. All rights reserved.

* Corresponding author. Tel.: +82 31 201 2514.

E-mail addresses: maqbool.hussain@oslab.khu.ac.kr (M. Hussain), muhammad.afzal@oslab.khu.ac.kr (M. Afzal), taqdir.ali@oslab.khu.ac.kr (T. Ali), rahmanali@oslab.khu.ac.kr (R. Ali), wajahat.ali@oslab.khu.ac.kr (W.A. Khan), arifj@skm.org.pk (A. Jamshed), sylee@oslab.khu.ac.kr (S. Lee), byeong.kang@utas.edu.au (B.H. Kang), khalid.latif@comsats.edu.pk (K. Latif).

<http://dx.doi.org/10.1016/j.artmed.2015.09.008>

0933-3657/© 2015 Elsevier B.V. All rights reserved.

1. Background

1.1. Motivations

Clinical decision support systems (CDSSs) play a pivotal role in improving patient care and enhancing practitioner performance [1]. Nevertheless, adaption of CDSSs in an actual healthcare workflow setup is challenging. Despite a long history of CDSS development, most of the systems evaluated in academia have not been realized in a real clinical practice environment. The most prominent barriers have included heterogeneous healthcare workflow integration, lack of standard knowledge representation, complexity of knowledge representation languages, lack of frameworks for clinical knowledge transformation into executable knowledge bases, and physician fears regarding validity of the services related to the knowledge bases and quality of published guidelines [2]. We therefore initiated the Smart CDSS project in collaboration with Shaikat Khanum Memorial Cancer Hospital and Research Centre (SKMCH)¹ to provide decision support services for head and neck cancer. We observed the above-mentioned barriers while working with clinical domain experts and hospital IT staff in gathering Smart CDSS requirements.

Notable problems in CDSS development that physicians face include the CDSS utility in their domain and the building of a model to acquire the knowledge in standard representation, such as Arden Syntax. To make the knowledge acquisition applicable, we developed data-driven knowledge acquisition and validation techniques, which use patient datasets for creation of the final clinical model, the refined clinical knowledge model (R-CKM). Data-driven knowledge acquisition helps in building a predictive model (PM) to reflect actual treatment decisions for patients recorded over the last ten years. Furthermore, the PM is validated against a clinical knowledge model (CKM) to finalize the R-CKM for the executable knowledge base. The CKM represents standard guidelines for cancer treatment published by National Comprehensive Cancer Network (NCCN) [3] <http://www.nccn.org/> (accessed 24.04.15). The R-CKM is transformed into medical logic modules (MLMs) using HL7 Arden Syntax. To avoid complexity of HL7 Arden Syntax, the MLMs are built in basic control artifacts, such as by using IF constructs. Moreover, to handle the intrinsic limitations of HL7 Arden Syntax, such as the curly brace problem [4] and binding issues with vocabulary [1], we adapt HL7 virtual medical record (vMR) [5] and use systematized nomenclature of medicine clinical terms (SNOMED CT) [6] <http://www.ihtsdo.org/snomed-ct/> (accessed 26.12.14) binding to represent all related concepts of head and neck cancer. The oral cavity is selected as the primary site for creation of R-CKM and the final candidate MLMs. The proposed approach is distinguished from available knowledge acquisition and validation approaches through its use of reference standard guidelines to validate the final knowledge model.

1.2. Related approaches

Knowledge acquisition and validation are prerequisites for effective decision support services. Various approaches are used to target objectives in target system design. In a methodological review, Peleg [7] categorized the approaches of translating clinical practice guidelines (CPGs) into computer-interpretable guidelines (CIGs): cognitive methods, a collaborative modeling methodology and tools, an information extraction methodology, and specialized CIG authoring tools.

Cognitive methods examine how people mentally represent information to solve subsequent problems. CPG translation to CIGs from narrative guidelines discussed in [8] is a cognitive method in which domain experts are provided with predefined algorithmic steps to develop the guidelines. Information extraction methods use semi-automatic translation by extracting knowledge using templates from narrative guidelines. Serban et al. [9] extracted templates from narrative guidelines and used them as building blocks for guidelines based on background thesaurus knowledge. CIG authoring tools are used to directly transform CPGs into executable CIGs. CIGs follow standard formats, such as XML, RDF, and any other standard format. Examples of such knowledge acquisition include use of an Arden Syntax editor to explicitly transform the clinical knowledge into an executable module, an approach mentioned in [10].

The systematic review [7] covers the methods that are based on translating narrative CPGs into CIGs; however, alternative approaches that use clinical datasets for knowledge extraction also exist. For example, Perera et al. [11] used a semantic-driven knowledge acquisition approach to establish missing relationships in data. Their proposed approach is limited to determining missing relationships in data whereby the initial knowledge base is constructed from UMLS vocabulary. Similarly, Gomoï et al. [12] used data mining techniques to generate rules from data and transform them into MLMs. This approach lacks a criterion definition for selection of candidate MLMs and final validation from reference guidelines. Our present study is closely related to [11,12] in terms of our use of clinical data as a common source of knowledge acquisition. In addition, we employ a cognitive method to align the refined model in accordance with published NCCN guidelines.

The study in [13] resembles our study with respect to the objectives of using a data-driven approach to create rules. The authors derive rules from patient data and incorporate them into guidelines for completeness of missing recommendations. The methodology proposed in the study comprises a rigorous inspection of guidelines to find missing recommendations for all possible patient symptoms. A decision tree learning algorithm is used to generate the rules, aligned with the guideline tree for missing recommendations. The proposed methodology is robust for the clinical domain with less complexity. For complex clinical domains with comprehensive guidelines, it may not be feasible to identify the missing decisions. Furthermore, as complexity of the clinical domain increases, the number of patient conditions is also increased to verify the final recommendation. With an increased number of patient conditions, it becomes difficult to derive related profiles (as used in this study) for which a set of rules needs to be generated. Finally, the study defines no formal validation criteria for the generated rules, and depends on the quality and quantity of data. In contrast, we use a data driven approach to select rules from the PM that conform to the guidelines. The PM is evolved to R-CKM using a rigorous validation process.

Most existing validation approaches aim to enable developed CIGs to capture the requirements of corresponding CPGs. These approaches can be viewed in two broad categories of techniques: inspection and testing [7]. In the inspection technique, domain experts investigate the CIGs for any possible errors in logic. Collaborative development of clinical guidelines is discussed in [14]. Teams of expert physicians and knowledge engineers create markups in final guidelines. These markups are evaluated together against the gold-standard markups with a thorough inspection to determine whether the final objective is achieved.

Testing techniques are used in addition to inspection techniques to minimize chances of errors in logic that may not be traced during inspection. These techniques use real patient cases to evaluate all possible decision paths in CIGs for obtaining correct decisions.

¹ SKMCH: <https://www.shaukatkhanum.org.pk/>

Miller [15] used a domain-constraint-based approach with a clinical condition set while generating the minimal set of test cases required for evaluating the particular guidelines. We herein employ both approaches: inspection and testing with different perspectives. Domain experts keep the PM decision path as a candidate path in R-CKM after validation (inspection) from the CKM. Knowledge engineers provide the evidence of patient cases from the PM that correctly classify it into a correct decision (testing). This validation process allows the decision path to be included in R-CKM if it conforms to the validation criteria (which is based on the CKM), which is provided with optimum accuracy from the patient data. Finally, R-CKM is validated against real patient data while transforming it into executable MLMs and integrated into hospital information system (HIS) workflows.

PMs are considered primary sources that physicians can use in clinical setups for recommendations. Widespread computational methods and tools are available for data analysis and creating of PMs. These methods and tools require a systematic method of selecting an appropriate PM that best fits the clinical prediction problem. The authors in [16] provided a systematic review of the most commonly used methods and simple guidelines for using these methods in clinical medicine. In our work, we use decision trees (DTs) to create the PM. We demonstrate CHAID and a classification and regression tree (CRT/CART) in our PM creation. Details on using basic PM techniques are found in [16–19].

Various standards for sharing clinical knowledge are available. HL7 Arden Syntax is one of the open standards of procedural representation of medical knowledge. The knowledge is represented in a modular logic unit – the so-called MLM – which is sharable across an organization [20]. HL7 Arden Syntax is used to specify a knowledge representation that is sharable, with the contents being readable by both humans and machines [1]. In this work, R-CKM, which was created from a PM, is represented in MLMs and evaluated in term of MLM performance on real patient medical data. We use HL7 Arden Syntax V2.7. However, readers can access detailed specifications with the current version (2.10) of HL7 Arden Syntax from the HL7 Arden Syntax working group repository [21].

Nevertheless, HL7 Arden Syntax has limitations in representing the ‘curly braces’ problem, the standard model used in logic. HL7 vMR is a standard data model proposed to resolve the curly braces problem that is associated with integration of MLMs with healthcare workflows [4,22]. The main intention of vMR is to create a simplified representation of a sufficient amount of clinical records for capturing information relevant to clinical knowledge and enabling understanding in knowledge engineers [5]. Most importantly, vMR is influenced by the HL7 V3 reference information model (RIM), which makes it easier to integrate it with HL7-compliant healthcare systems. The HL7 vMR model was developed based on analysis of requirements from 22 US institutions [22]. We create a cancer clinical domain model derived from HL7 vMR and used as a data model for creating MLMs. To this end, we employ HL7 vMR version 1; however, readers may consult the recent version 2 of HL7 vMR for detailed specifications [5].

2. Methods

Clinical knowledge acquisition is the main activity in achieving successful deployment of CDSSs. Unlike conventional requirements gathering and modeling of a system, knowledge creation needs a detailed set of activities that cover the actual scenarios and facts occurring in a real environment. Most importantly, the domain experts are not required to know the executable knowledge paradigm used as an integral part in a real healthcare workflow. Moreover, the ultimate goal of knowledge acquisition is to represent the knowledge that functions with an existing healthcare workflow and to enable a proper validation process for the final knowledge model. In this regard, we adapt a data-driven knowledge acquisition and validation approach that reflects a real clinical setup deriving the clinical knowledge from existing clinical practices. The proposed approach is an iterative model that includes three phases with ten activities. It enables a domain expert to create an executable knowledge base with coordination of a knowledge engineer. An abstract view of the phases with activity descriptions is depicted in Fig. 1.

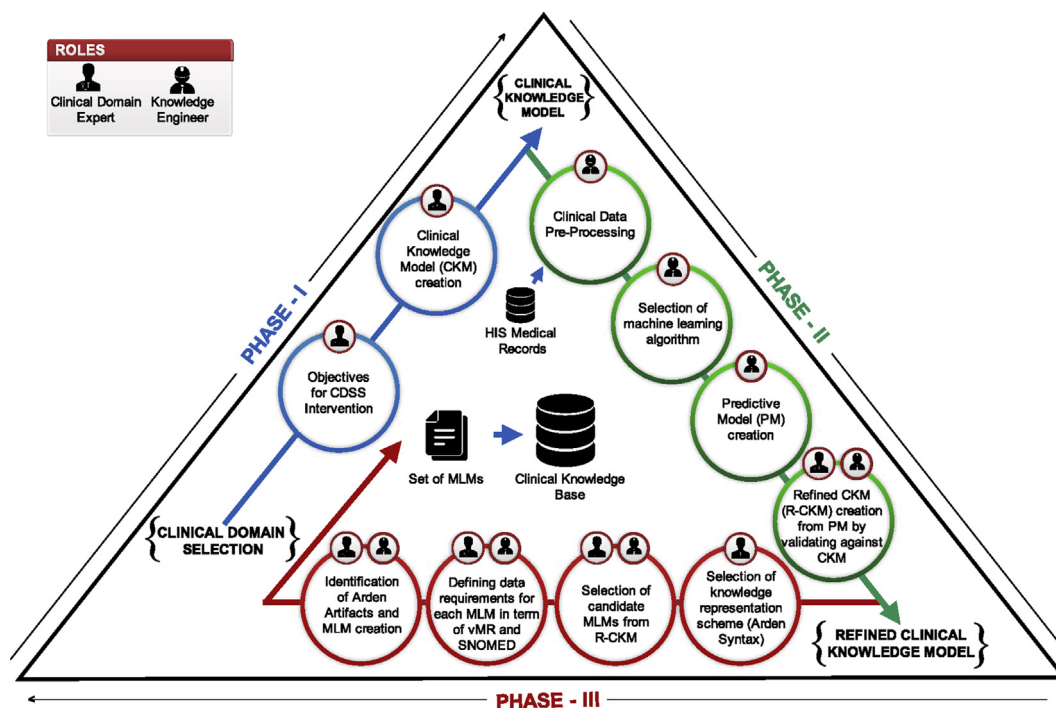


Fig. 1. Data-driven clinical knowledge acquisition and validation process model.

2.1. Phase-I: clinical knowledge modelling

A CKM is the baseline model for validation of ultimate clinical rules in a knowledge base. It represents the clinical knowledge for a specific clinical domain, while referencing standard knowledge resources, such as CPGs. CPGs are widely used consulting knowledge resources that are applied in clinical practices for diagnoses and treatments. CPGs are available in different formats, such as textual narratives and/or decision trees. A CKM is a formal decision tree representation of CPGs aligned with clinical objectives in a particular domain. In Phase I, a team of physicians is involved to finalize the CKM and perform the following activities.

Defining a set of clinical objectives: Clinical objectives specify the scope and intended outcomes of knowledge acquisition. In the context of a CDSS, clinical objectives are defined to specify possible CDSS interventions.

- We define the clinical objective for creating clinical knowledge covering a treatment plan recommendation for head and neck cancer. The treatment plan includes a single procedure or combination of procedures from radiotherapy, chemotherapy, induction chemotherapy, and surgery. The clinical knowledge helps physicians during treatment of a patient or gives recommendations of a treatment plan during a multidisciplinary conference² of physicians.
- The recommendations are classified based on a particular tumor site. In this paper, the oral cavity is considered a primary site for an initial CDSS recommendation.

Selection of CPGs: Physicians consult CPGs for diagnoses and treatment plans during medical practices. In this step, physicians intend to select appropriate CPGs that align with the clinical objectives and use an appropriate knowledge representation scheme. In current work, SKMCH physicians use the following guidelines:

- NCCN guidelines are used as candidate CPGs to model clinical knowledge for tumors of the oral cavity and other sites.
- Tumor, Node, Metastases (TNM) staging guidelines are used to represent the clinical staging of tumors.
- A decision tree is selected as a formal representation of knowledge. Decision nodes are represented as rectangles and diamonds, while the conclusion nodes are represented with oval corner rectangles. Oval corner rectangles also play the role of condition nodes in case it comes in the middle of the tree.

Creation of CKM: A team of physicians converts the selected guidelines into a formal decision tree representation. For oral cavity CKM development, the following steps are used:

- Two resident doctors are assigned to initially create the draft decision tree from NCCN guideline trees and narratives.
- The initial draft of the decision tree is thoroughly inspected by a senior oncologist and approved as a final CKM with a possible amendment if needed.
- In the final CKM decision tree, a senior oncologist may incorporate some proven practices that may not be included in CPGs but bear evidence of its validity from other knowledge.

2.2. Phase-II: knowledge acquisition and validation

Knowledge acquisition and validation are a core aspect of this work. They are achieved through application of machine learning

algorithms on existing medical records in HIS and applying a rigorous validation process supported by the CKM. Phase II comprises activities that are categorized into two broad perspectives: data-driven knowledge acquisition and knowledge validation.

2.2.1. Data-driven knowledge acquisition

Patient medical records in HIS are the primary resource for acquisition of clinical knowledge. Patient medical records reflect patient encounters and detail histories of diagnoses and ongoing or completed treatment. Various computational methods and tools are widely used for analyzing patient medical records and creating PMs for future recommendations. The authors in [16] discussed overall issues and provided state-of-the-art guidelines to use appropriate methods for PM creation in clinical medicine. In this study, patient data of head and neck cancers with tumors in the primary oral cavity site were imported into SPSS [19] <http://www.sussex.ac.uk/its/pdfs/SPSS.Decision.Trees.21.pdf> (accessed 24.12.14) and Weka [23] from HIS to create a PM using a decision tree (DT) classification method. The main advantage of DTs is the comprehensibility of the classification structure, whereby they can easily determine attributes for classifying and verifying new data [18]. In addition, owing to powerful heuristics, the computational complexity of the DT induction algorithm is low [16]. Finally, it provides the opportunity to generate readily comprehensible knowledge rules.

In the context of tasks and guidelines provided by [16] for the PM, the data-driven knowledge acquisition takes into account the following related activities. Fig. 2 highlights the detailed tasks in each activity and possible sequences for performing these activities.

Clinical data description and pre-processing: The clinical data of oral cavity cancer patients for this study was imported from HIS with thirteen condition attributes and one decision attribute (treatment plan). Details of the condition attributes and decision attributes are shown in Table 1. For the final prediction model, 1229 patient records were used after applying pre-processing (removing and calculating missing attributes) on 2181 original patient records. Detailed steps of the pre-processing are shown in the “Data preprocessing” activity in Fig. 2.

Selection of machine learning algorithm: The main goal of this activity is to determine the appropriate best-performing decision tree algorithm on a given dataset for generating the final PM. PMs can be evaluated based on their predictive performance and comprehensibility. Predictive performance can be quantified using classification accuracy, while comprehensibility is a subjective measure that is assessed by a domain expert. In our context, we combine both measures into quantitative measures to achieve our desired criteria. The criteria defined by the domain expert is the generating of a PM with high accuracy and providing a minimal set of decision paths by involving fewer dominant condition attributes. This criteria is translated into a quantitative measure using the weighted sum model (WSM). WSM ranking is expressed in Eq. (1), which uses classification accuracy P , the number of rules generated R , and the number of attributes A involved in the conditions. Weights w_j are assigned based on the importance of the attributes. Classification accuracy P is the most important in the selection of algorithm, which is assigned $w_p: +0.8$. The number of rules R and number of attributes A are assigned $w_r: -0.1$ and $w_a: -0.1$, respectively. According to our criteria, an algorithm with a minimum rule set and involving fewer attributes is preferred. In this regard, we choose a negative scale to discourage algorithms that generate maximum rules and/or those involving more attributes.

According to our criteria, CHAID is the most suitable algorithm among CRT, QUEST, DFTree, and J48 for use with the PM and rules

² A multidisciplinary conference comprises a panel of doctors including oncologists, radiologists, surgeons and other resident physicians. They conduct a conference on regular basis to select the final treatment plan for a patient.

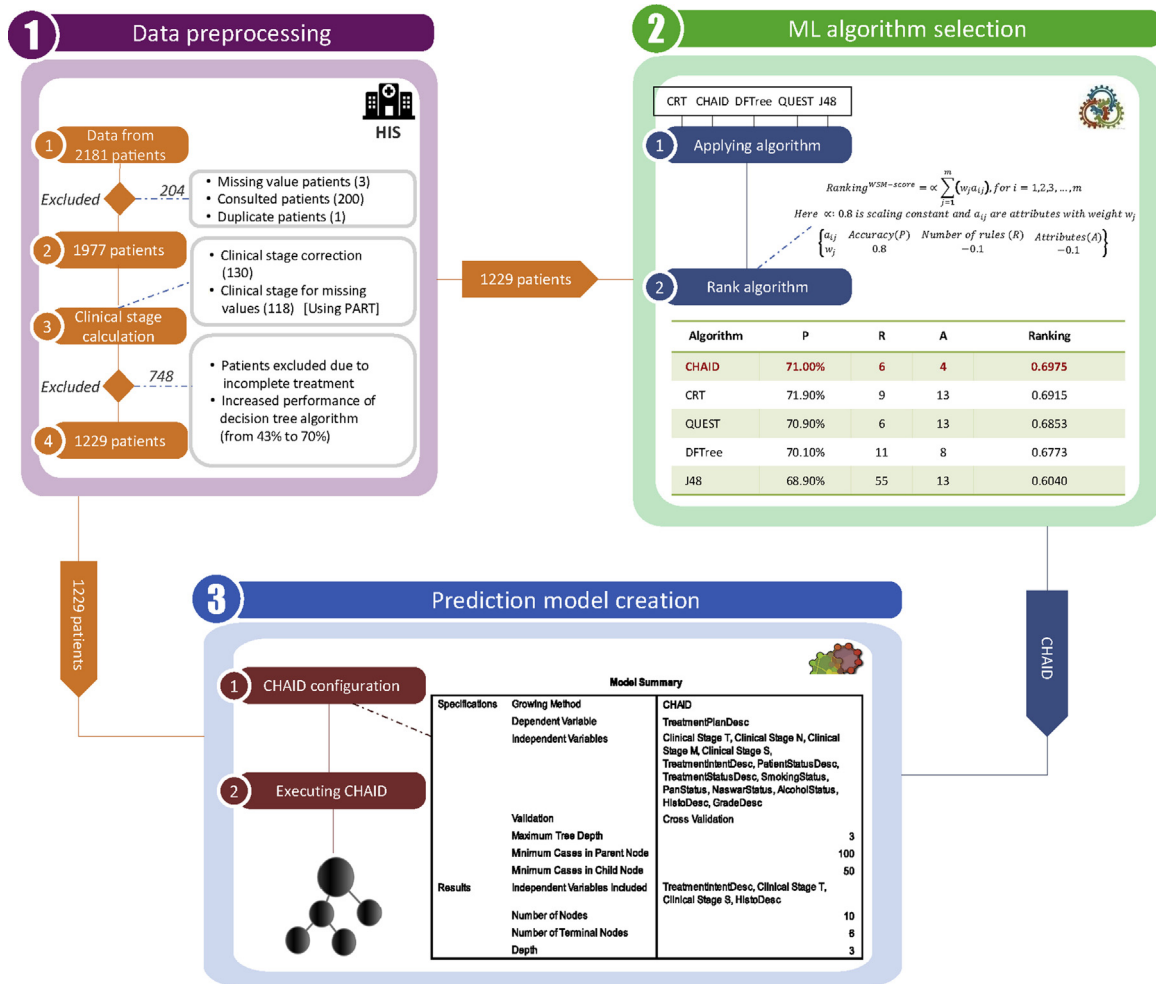


Fig. 2. PM creation process.

generation, as shown in the “ML algorithm selection” activity in Fig. 2.

$$\text{Ranking}^{\text{WSM-score}} = \alpha \sum_{j=1}^m w_j a_{ij}, \quad \text{for } i = 1, 2, 3, \dots, m \quad (1)$$

Here α : (0.8) is scaling constant and a_{ij} are attributes with weight w_j

$$\left\{ \begin{array}{llll} a_{ij} & \text{Accuracy}(P) & \text{Number of Rules}(R) & \text{Attributes}(A) \\ w_j & 0.8 & -0.1 & -0.1 \end{array} \right\}$$

Creation of a PM: CHAID is an appropriate candidate algorithm that provides a PM with reasonable classification accuracy on given data while generating easily understandable rules.

CHAID uses multiway splits to generate more than two nodes from a current node. It chooses the independent (predictor or condition attribute) variable with the strongest interaction with the dependent variable (decision attribute). It has the capability of merging the category of each predictor if they are not significantly different with respect to the dependent variable. A detailed description of CHAID and other tree algorithms are given in [18,19].

CHAID is applied on a 1229-patient dataset using the 13 condition attributes mentioned in Table 1. The algorithm has default

parameters of the SPSS tool, as shown in the “prediction model creation” activity in Fig. 2.

A decision tree representation of the PM is shown in Fig. 3. The confusion matrix shown in Table 2 presents the overall accuracy of the model. In summary, our final model achieved 71.0% accuracy for classification of final cancer treatment.

The PM of CHAID classified the designated treatment plan with the accuracy of C CRT (85.5%), S RT (80.5%) and RT (46.2%). The accuracy of RT was comparatively low, but the major proportion of cases (127;59.4%) of remaining cases (214) were classified as S RT. The S RT treatment plan covered radiotherapy (RT) following surgery as the main procedure; therefore, it compensated the lesser precision of the classification model to only the RT decision class. There are many reasons for the direct RT treatment rather than following standard treatment of S followed by RT. These include patient not willing for surgery or having some comorbidities associated with tumor site. The CHAID classification algorithm has the intrinsic property of selecting dominant attributes from a set of condition attributes, which provided a higher segmentation of data. It selected 4/13 attributes “Treatment Intent Description”, “Clinical Stage T”, “Clinical Stage S” and “Histology Description”. The model generated ten nodes overall, of which six nodes were terminal (leaf) nodes in the tree.

2.2.2. Knowledge validation

Knowledge validation is performed on the PM with conformance criteria that ensure that the final R-CKM model conforms

Table 1
Data description of oral cavity.

Data summary: Total attributes: 14 Decision attribute: treatment plan description		
Attribute type	Attribute title	Attribute description
Decision attribute	Treatment plan description	Treatment plan for patient. C: Induction chemotherapy is done. CRT: Chemoradiotherapy is done. RT: Radiotherapy is done. S: Surgery is done. C CRT: C is done followed by CRT. C RT: C is done followed by RT. S CRT: S is done followed by CRT. S RT: S is done followed by RT. C S CRT: C is done followed by S and CRT. C S RT: C is done followed by S and RT. UK: Treatment unknown. NA: Treatment not applicable.
Condition attributes	Grade description Treatment intent description Treatment status description Clinical Stage T Clinical Stage N Clinical Stage M Clinical Stage S Smoking Alcohol Naswar Pan Patient status Histology description	Indicate patient treatment status such as poor and moderate. Patient status for treatment, such as palliative or radical. Indicate patient treatment status such as completed. TNM Staging T value. TNM Staging N value. TNM Staging M value. TNM Staging S value. Smoking status. Alcohol status. Naswar status. Naswar is a moist, powdered tobacco snuff. Pan status. Pan is type of, tobacco chewed and finally spat out or swallowed. Patient current status such as alive and dead. Indicate patient disease such as Carcinoma.

to the CKM (standard CPGs). Domain experts and knowledge engineers worked in a collaborative environment to validate the PM against CKM using well-defined inspection and testing mechanisms and developed the R-CKM conformed and validated model. Fig. 4 depicts the process of validation with detailed flows of steps.

Activities performed in the validation process can be classified into three main categories: setting the validation criteria, validating the PM against the validation criteria, and finally evolving the R-CKM by inspecting and refining the PM.

Setting validation criteria: The validation criteria are a set of assertions that may be required to pass the decision paths in PM to be eligible for inclusion in R-CKM. Domain experts specify the validation criteria, which is influenced by their practices and conforms to the evidence from standard guidelines (CKM). While specifying the criteria, each criterion is assigned a priority and its primary status. The priority dictates the order of execution in the validation process; the primary status specifies that the given criteria must be satisfied by the decision path. Table 3 provides the set of criteria defined by domain experts to validate the oral cavity PM in this work.

PM validation against criteria: Validation is an iterative process that selects one decision path P_i at a time from the PM for validation. The decision path is selected as part of R-CKM if it satisfies

all primary validation criteria and passes at least one of the non-primary criteria. For example, any decision path in the oral cavity PM becomes part of R-CKM if it satisfies criteria 1 and 2 and either 3 or 4, as mentioned in Table 3. The iteration is finished once all the decision paths in the PM are evaluated against the validation criteria.

Inspection and refinement of selected PM decision path: The decision path P_i in the PM that passes the validation criteria is inspected and refined to P_j prior to becoming part of R-CKM. Inspection and refinement involves activities to identify conditional and decision values in a decision path for the same interpretation with existing CKM conditional and decision values. Therefore, in the refined decision path, the concepts are aligned according to the CKM. For example, the conditional value clinical stage S: 1 is interpreted as clinical stage T: 1 and clinical stage N: 0 in the CKM. Similarly, during refinement, the physician may add some other treatment to already given choices by providing evidence for the inclusion.

2.3. Phase-III: R-CKM transformation into executable rules

Most of the projects involving data mining and machine learning techniques in the clinical setup are stopped after the PM is created. This is unfortunate because the PM should be deployed in a real clinical setup for clinical decision support as a daily base service.

Table 2
Classification using the CHAID model for oral cavity treatment planning.

Observed	Predicted			Total cases	Accuracy
	C CRT	RT	S RT		
C CRT	343	12	46	401	85.5%
RT	87	184	127	398	46.2%
S RT	81	3	346	430	80.5%
Total	511	199	519	1229	
Overall percentage	41.6%	16.2%	42.2%		71.0%

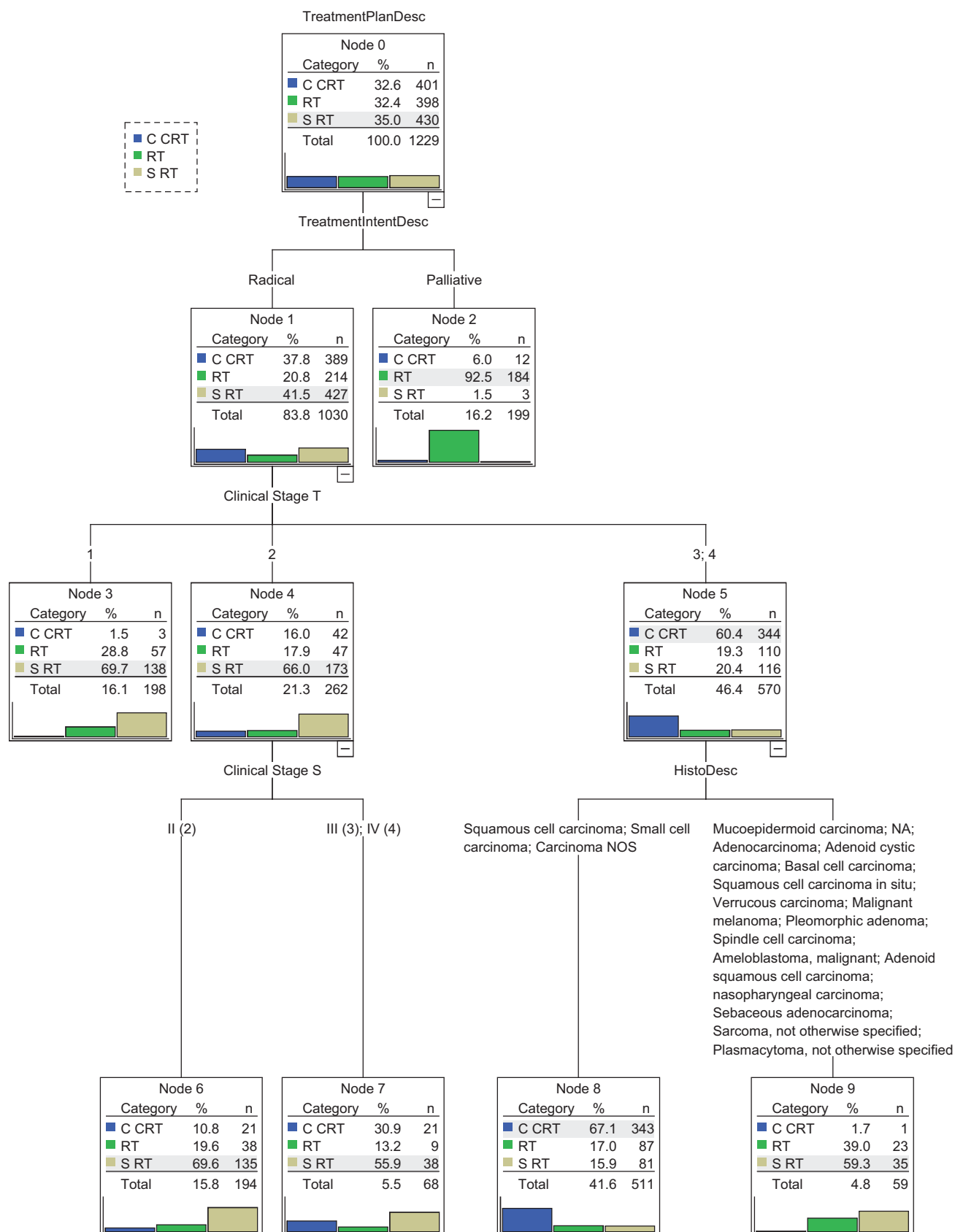


Fig. 3. CHAID prediction model for oral cavity treatment plans.

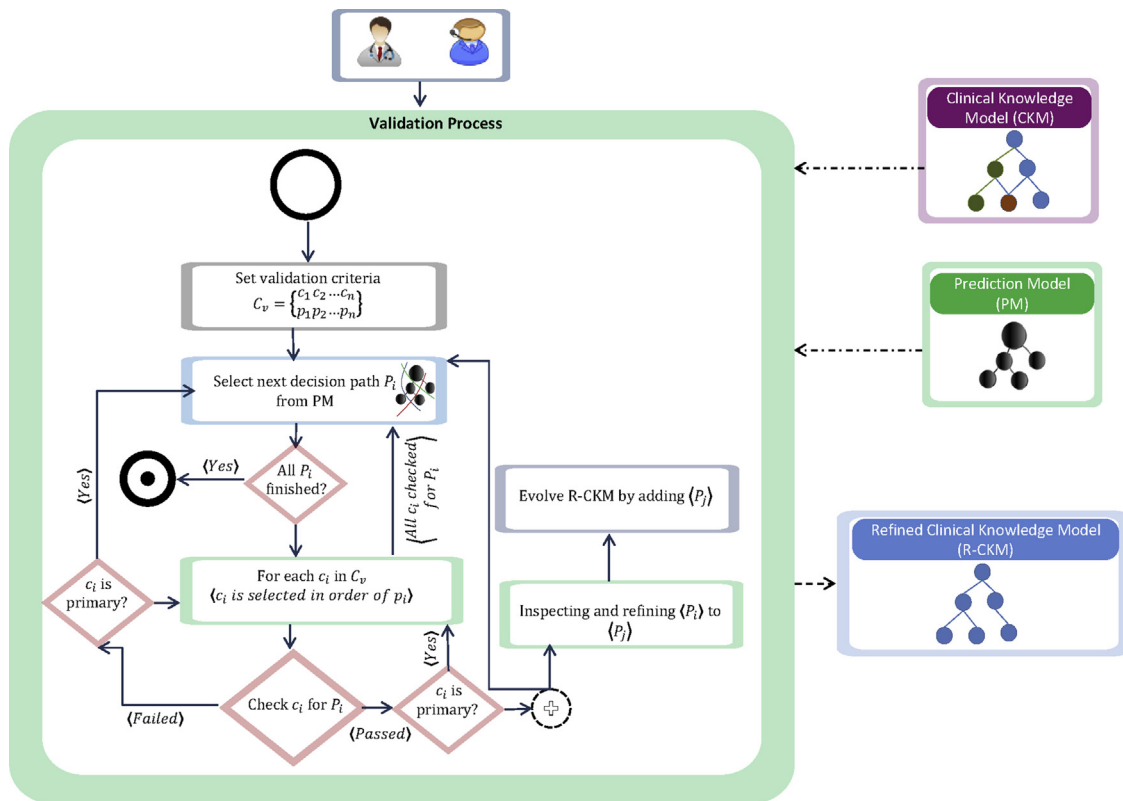


Fig. 4. Knowledge validation process.

This would help in supporting the medical services, and stakeholders could easily monitor the efficiency of the technology in terms of improving quality of care and decreasing healthcare costs [16]. Maintaining the motivation to provide end-to-end implementation of the PM, R-CKM is represented in sharable format using HL7 Arden Syntax. R-CKM is converted into a set of candidate MLMs and implemented in a real setup for recommendation of a treatment plan for oral cavity cancer patients. Conversion of R-CKM into executable knowledge representation (MLM) is performed using the following activities.

Selection of candidate MLMs from the R-CKM: R-CKM can be transformed into different sets of MLMs depending on the domain expert intuitions and logical connections existing in the decision path of R-CKM. For R-CKM, three candidate approaches are

analyzed for final executable knowledge. These are explained in Section 3.3.

Data requirements for MLMs using the HL7 vMR model: Data specifications for each MLM are important for the formal creation of logic. Data specifications include enlisting clinical data that is required for the MLM, representing clinical data in the standard data model (HL7 vMR), and mapping coded concepts into a standard vocabulary. Section 3.3 details the specifications of data requirements for candidate MLMs.

Identification of HL7 Arden Syntax artifacts and MLMs creation:

Arden Syntax is a comprehensive specification supporting a large number of operators, various control structures, including decision and looping structures, and comprehensive models for various data types. Knowledge engineers summarize the basic artifacts required to transform the R-CKM into corresponding

Table 3
Validation criteria for oral cavity.

C. no.	Criteria	Priority	Primary	Remarks
1.	$\{ \forall P_i \in PM : Accuracy(P_i) > N\% \}$	1	Yes	<ul style="list-style-type: none"> The domain expert assigns N, which represents the accuracy of PM based on the training data. Trade off: Higher accuracy setting produces an efficient model, but coverage of involving more patient features is limited and vice versa.
2.	$\{ \forall P_i \in PM \wedge \forall P_j \in CKM : ! Conflict(P_i, P_j) \}$	1	Yes	<ul style="list-style-type: none"> Conflicts with guidelines; conflicting treatments must not be exist. Example: after surgery, chemo-induction has no meaning.
3.	$\{ \forall P_i \in PM \wedge \exists P_j \in CKM : Conform(P_i, P_j) \xrightarrow{yields} P_i \in \Delta RCKM \}$	2	No	<ul style="list-style-type: none"> Decision path in PM conforming to any CKM path shall be part of R-CKM.
4.	$\{ \exists P_i \in PM \wedge \forall P_j \in CKM : !Conform(P_i, P_j) \xrightarrow{provides} Evidence(P_i) \xrightarrow{yields} P_i \in \Delta RCKM \}$	3	No	<ul style="list-style-type: none"> Decision path in PM not conforming to any path in CKM can be part of R-CKM only if: Sufficient evidence exists for effectiveness of the treatment. Evidence can be other standard clinical knowledge resources or local practices with a reasonable success ratio for the predicted treatment.

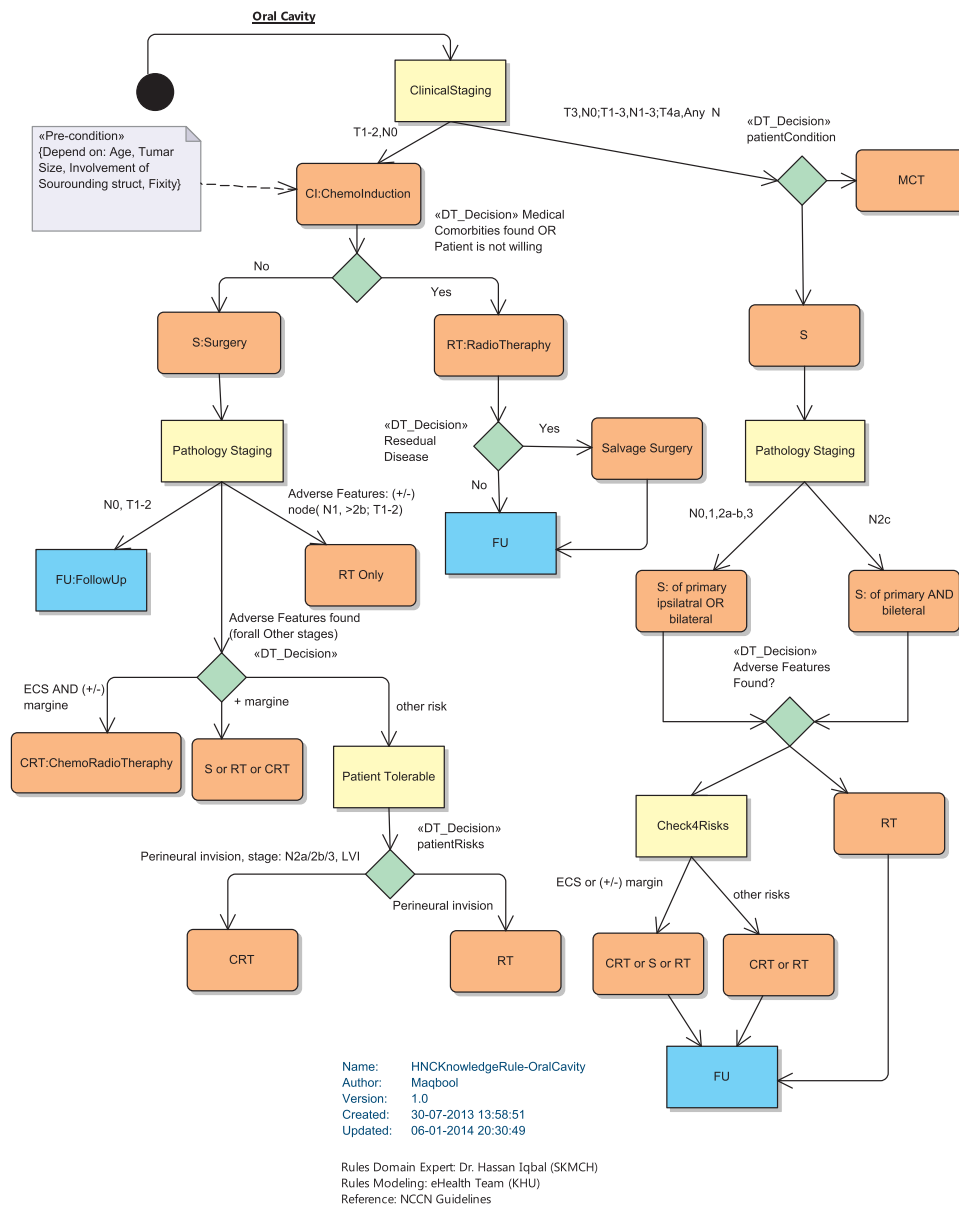


Fig. 5. Clinical knowledge model for the oral cavity using the NCCN treatment plan.

MLMs and provide physicians with training on using these artifacts.

Integration of MLMs with HIS workflow: Knowledge engineers implement the MLMs and integrate them with HIS workflows. Details of the implementation and integration of MLMs with HIS are available in our previous research [24,25].

3. Results

3.1. Phase-I: clinical knowledge modelling

The team of physicians establish the clinical objective of incorporating the CDSS intervention for the *treatment plan recommendations for patients with an oral cavity tumor*. NCCN guidelines are selected to develop the CKM for the decision of treatment plans for these patients. Fig. 5, shows the final outcome of the CKM of the clinical knowledge modeling phase from the NCCN guidelines. These guidelines cover the domain as a whole and provide a general view of the decision model. In order to convert NCCN into the final

CKM decision tree model, physicians also integrate local practices into decision paths with sufficient evidence for improving patient care. For example, *ChemoInduction* was added to the final CKM with local practices, suggesting that induction chemotherapy followed by chemoradiotherapy before surgery improves overall survival of patients [26].

3.2. Phase-II: knowledge acquisition and validation

PM evaluation using disjoint patient test data: The experiment is performed on a (disjoint) dataset of 739 patients with completed treatments. Each decision path of the PM is evaluated for a set of candidate patient cases. The patient cases are distributed into six decision paths, where distribution is based on qualification for the condition part in the path.

The overall accuracy of the PM is calculated as the weighted mean of the accuracies of all of the decision paths, as shown in Eq. (2). PM accuracy is 59.0% in the test data, which is encouraging to consider the model for constructing the final R-CKM. Table 4 shows

Table 4
PM evaluation on disjoint patient test data.

Path#	PM decision path	Candidate patient cases	PM path accuracy
Path-1	Node0 → RT	Palliative patients: 69	40.58% (C:28 W:41)
Path-2	Node0 → Node1 → Node3 → S RT	Patient with radical and CS: T1: 139	95.68% (C:133, W:6)
Path-3	Node0 → Node1 → Node4 → Node6 → S RT	Patients with radical and FCS II: 123	73.98% (C:91, W:32)
Path-4	Node0 → Node1 → Node4 → Node7 → S RT	Patients with radical and FCS III/IV: 56	67.86% (C:38, W:18)
Path-5	Node0 → Node1 → Node5 → Node8 → C CRT	Patients with radical, CS:T3-4 and HistoDescription 1, 2, 3: 324	38.58% (C:125, W:199)
Path-6	Node0 → Node1 → Node5 → Node9 → S RT	Patients with radical, CS:T3-4 and HistoDescription other than 1, 2, 3: 28	85.71% (C:24, W:4)
Overall PM Accuracy: PM_{acc}			59.0%

- CS: Clinical Stage.
- FCS: Final Clinical Stage.
- HistoDescription 1: Squamous cell carcinoma.
- HistoDescription 2: Small cell carcinoma.
- HistoDescription 3: Carcinoma NOS.
- C: Correctly classified patient cases.
- W: Wrongly classified patient cases.

the detailed distribution of patient test cases over six decision paths and their corresponding accuracies. The “PM Decision Path” represents the decision paths covered by PM, “Candidate Patient Cases” are the record number of cases that qualify the decision path and “PM Path Accuracy” is the percentage of patient cases correctly classified by the PM decision path.

$$PM_{acc} = \frac{\sum_{i=1}^n (pat_{c_i} \times A_{dp_i})}{\sum_{i=1}^n pat_{c_i}} \quad (2)$$

where pat_{c_i} and A_{dp_i} represent the number of patient cases assigned to path dp_i and accuracy of path dp_i respectively.

Validated R-CKM based on PM: R-CKM is created from the PM while using the validation criteria defined by domain experts. Using a set of four validation criteria for the oral cavity, the validation process is applied on the oral cavity PM, which resulted in R-CKM, as shown in Fig. 6.

Table 5 shows the details of the applicable validation criteria (AVC) for each decision path of the PM. For validation criteria 1, physicians establish the accuracy level of $N = 50\%$ (accuracy of paths based on training data). Five decision paths in the PM are conformed to the CKM (i.e., passing validation criteria of 1, 2, 3), while one decision path represent the local practices (passing criteria 1, 2, 4). Moreover, the PM decision always existed in leaf nodes, whereas in R-CKM the decision node could have occurred in the middle and would not work as a conditional node for the following sub-tree. For example, in the PM, decision

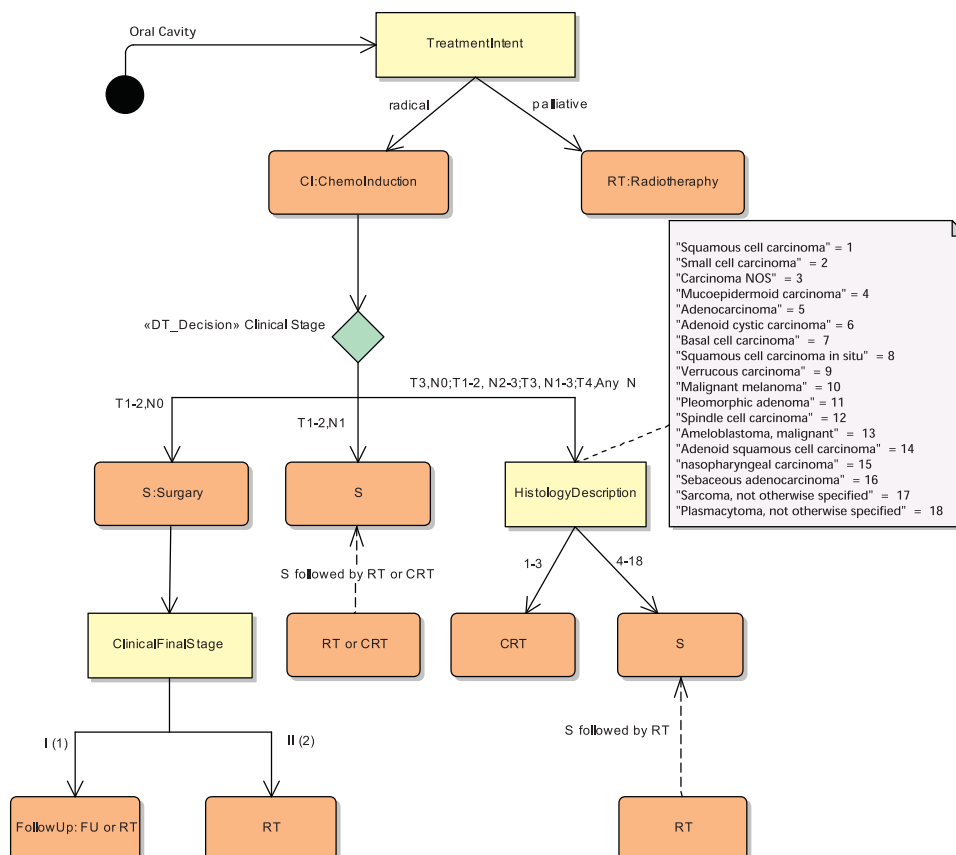


Fig. 6. Refined clinical knowledge model for the oral cavity treatment plan

Table 5
R-CKM evolution details.

Decision path (PM)	AVC	Refined decision path (R-CKM)	Remarks (using CKM)
N0 → N2	{1, 2, 3}	TreatmentIntent → RT	<ul style="list-style-type: none">Conforms to CKM as:<ul style="list-style-type: none">RT is a secondary-level treatment for clinical stage III and IV patients in CKMPalliative patients have stage III or stage IV (HIS statistics: 90%)Thus, RT for palliative care conforms
N0 → N1 → N3	{1, 2, 3}	i. TreatmentIntent → CI → S ii. TreatmentIntent → CI → S → FU or RT	<ul style="list-style-type: none">Conforms to CKM as:<ul style="list-style-type: none">Surgery is done followed by RT for all clinical stages in CKMRadical patients may have any of the clinical stages (HIS statistics: I: 17.48%, II: 18.83%, III: 17.67% and IV: 46.02%)Thus, the given decision path conforms to CKMFU: FollowUp is added to refine the path because after pathology stage I some patients may ask for FU
N0 → N1 → N4 → N6	{1, 2, 3}	i. TreatmentIntent → CI → S ii. TreatmentIntent → CI → S → RT	<ul style="list-style-type: none">Same conformance as for N0 → N1 → N3
N0 → N1 → N4 → N7	{1, 2, 3}	TreatmentIntent → CI → S → RT or CRT	<ul style="list-style-type: none">Same conformance as for N0 → N1 → N3CRT is added to refined path because N7 in PM suggests CRT with 30% accuracyMoreover, CRT is a tertiary level treatment in CKM for advanced clinical staging
N0 → N1 → N5 → N8	{1, 2, 4}	TreatmentIntent → CI → CRT	<ul style="list-style-type: none">Not conforming to CKM, but physicians provide the following evidence from existing practicesC CRT makes 32.63% of the dataset and 84.5% patients among C CRT are stage III and stage IVBecause PM shows significant accuracy, 67.1%, for C CRT and is very effective
N0 → N1 → N5 → N9	{1, 2, 3}	TreatmentIntent → CI → S → RT	<ul style="list-style-type: none">Same conformance as for N0 → N1 → N3

path N0 → N1 → N4 → N6, N6 is the decision node representing S RT (surgery followed by radiotherapy). The same decision path is divided into two decision paths in R-CKM; i.e., TreatmentIntent → CI → S and TreatmentIntent → CI → S → RT, where S (as the decision node) in the first decision path is the condition node for the second decision path.

3.3. Phase-III: R-CKM transformation into executable rules

Evaluation of R-CKM to MLM strategies: Clinical models can be transformed into different sets of MLMs depending on the domain expert intuitions and logical connections in the decision path. For R-CKM, three candidate approaches were analyzed for the final executable knowledge.

The creation of a single MLM with a single event, which covers all decision paths of the R-CKM, is the most common approach. This approach has the advantage of publishing minimal MLMs to executable clinical knowledge, which makes it easy to maintain the growing knowledge base. However, this approach has limited MLM re-usability. Furthermore, it becomes difficult to identify and fix logical errors in a single MLM depending on a large clinical model.

The second approach is to create a set of MLMs for a given R-CKM shown in Fig. 6; that is, create an individual MLM with a common event for each single decision node or a set of decision nodes sharing the immediate decision node. For R-CKM, six candidate MLMs can be created: one MLM for tree level 1 (covering ChemoInduction and Radiotherapy), one for tree level 2 (covering Surgery), two for tree level 3 (MLM covering RT or CRT and MLM covering CRT and S), and two for tree level 4 (MLM covering FollowUp or RT and RT and MLM covering RT). In this case, MLM for children decision node(s) include logic for a complete selected path originating from the root. For example, the MLM at level 2 concludes with Surgery after checking that the patient has done ChemoInduction and (s)he was on a radical treatment care. This approach has the potential to generate separate MLMs

for each decision node, which can easily be traced to the original clinical model without digging into the detailed logic. Moreover, this approach is close to the rule generation of most machine learning tree classification, such as CHAID. The main limitations of this approach include duplications due to logic tracing parent nodes, multiple MLMs evoked for a single event, which will need the same amount of data from HIS, and maintenance problems occurring in the case a change is made to the clinical model.

The third approach is to create a set of MLMs controlled by the root MLM through the MLM-calling mechanism. It allows modular logic to create reusable and understandable MLMs. The root MLM is exposed to a trigger for a particular event and imports all data required for the decision logic used in the “called” MLMs – so called *subMLMs*. SubMLMs can also play the role of “caller” for other subMLMs. Using this approach, a different set of MLMs for the same clinical model can be created depending on the modularity of the logic. For R-CKM, four candidate MLMs are selected to cover all decision paths. Fig. 7 shows a detailed decision part of each MLM highlighted in R-CKM.

At first level of R-CKM, RootMLM:OralCavity_LocalizedPrimitive is selected as the primitive root MLM, which is exposed for the oral cavity treatment plan event. It assigns a decision part at level 1 of the model for the primitive treatment plan. For further decisions it calls subMLM:OralCavityComplexLv11 MLM, which represents the logic at level 2 based on clinical staging. Some treatment plans at level 2 and all other treatment plans in subsequent levels (i.e., levels 3 and 4) are delegated to two subMLMs OralCavityComplexLv121 and OralCavityComplexLv122, which are called from subMLM:OralCavityComplexLv11. RootMLM:OralCavity_LocalizedPrimitive is the controller MLM triggered on the oral cavity treatment plan, and recommendations are made in coordination with subMLMs. This approach avoids duplication in logic. Furthermore, distribution of logic in subMLMs gives control to managing the overall logic of the clinical model.

Data specifications for MLMs and mapping to a standard model (HL7 vMR): Data specifications for each MLM are important for formal development of internal logic. Data specifications

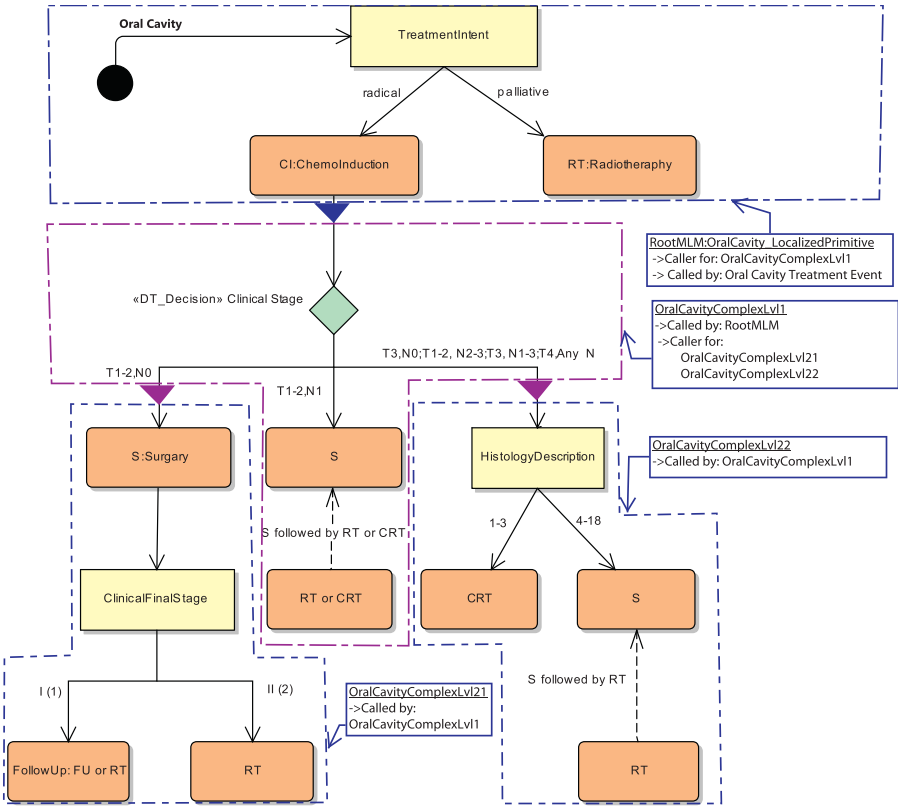


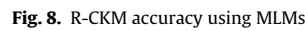
Fig. 7. Candidate MLMs for oral cavity R-CKM

include activities, enlisting clinical data required for the MLM, representation of clinical data in a standard data model, and mapping of coded concepts into a standard vocabulary. For the oral cavity treatment event, four candidate MLMs need different sets of clinical data to recommend an appropriate treatment plan. OralCavity.LocalizedPrimitive is the root MLM triggered against the oral cavity treatment event; therefore, the data required for subMLMs are imported by the root MLM and passed as an argument to caller subMLMs. The root MLM OralCavity.LocalizedPrimitive requires data regarding treatment intent and primary level treatment for the treatment plan recommendation at the primary level. It needs further detailed data if primary level treatment is done and it delegates control to subMLM with corresponding data for secondary or tertiary level recommendations. OralCavityComplexLv1 subMLM is called by the root MLM OralCavity.LocalizedPrimitive and expects clinical data regarding clinical staging T and N values and treatment completed at

secondary and tertiary levels. It also expects clinical data regarding clinical staging S and disease (histology findings), which are passed to subMLMs OralCavityComplexLv21 and OralCavityComplexLv22, respectively, if necessary to call further recommendations. Clinical data specified for each MLM are required to have standard representation of vMR that can be understood by CDSS. All individual clinical data concepts are needed for mapping into vMR concepts. Table 6 lists HIS data mapping to vMR concepts in detail. A detailed clinical model for oral cavity treatment used by candidate MLMs is shown in Appendix A. As indicated in Table 6, all of the values used in clinical data are coded concepts and mapped to vMR coded attributes. SNOMED codes for all HIS concepts and the corresponding value set is searched in the SNOMED repository. Table B.8 (Appendix B) lists all related SNOMED codes associated with HIS concepts and corresponding values.

Table 6
HIS clinical data mapping to HL7 vMR.

HIS concepts	vMR concepts	Attributes mappings (HIS-vMR)
Clinical Stage T	ObservationResult	Clinical Stage T = observationFocus Clinical Stage T value = observationValue
Clinical Stage N	ObservationResult	Clinical Stage N = observationFocus Clinical Stage N value = observationValue
Clinical Stage S	ObservationResult	Clinical Stage S = observationFocus Clinical Stage S value = observationValue
Treatment intent	ProcedureEvent	Treatment intent = procedureCode Treatment intent value = procedureMethod
Histology description	Problem	Histology description = problemCode
Treatment plan	ProcedureEvent	Treatment plan = procedureCode Treatment plan value = procedureMethod



Please cite this article in press as: Hussain M, et al. Data-driven knowledge acquisition, validation, and transformation into HL7 Arden Syntax. *Artif Intell Med* (2015), <http://dx.doi.org/10.1016/j.artmed.2015.09.008>

mentioned in the guidelines and recorded attributes in healthcare system. From the current study of oral cavity CKM and the available practices dataset, it is indicated that clinical staging is common. On the other hand, pathology staging and other adverse conditions from guidelines are not recorded as part of the patient health record. Similarly, the most notable risks (such as history of smoking, pan and naswar) and other problems and patient conditions (such as disease information, patient status) do not exist in the CKM. In this context, PM plays an important role in generating rules that involve most dominant attributes from data for the final decision. Creation of R-CKM from the PM with proper validation from the CKM produces a model that can be directly evoked with existing data and conforms to the standard guidelines. As shown in Table 5, all decision paths except one from PM conform to the CKM and are represented in R-CKM. Inclusion of the CRT recommendation is based on strong evidence from classification of high proportions of patient cases for clinical stage III (3) and IV (4) and on confirmation of its effectiveness in patient care by the physicians.

4.2. R-CKM validation vs verification

Validation is intended to ensure the building of the correct model, while verification ensures that the built model is working correctly. After validation, the CIGs are checked for internal consistencies using some formal mathematical models, so-called verification [7]. Different approaches are available for verifications. A model-driven approach creates a model of specifications and verifies it against temporal logic [27,28]. A theorem-proving approach uses formal methods to represent the guidelines in a formal model and proves the model for the verification of interesting properties of the guidelines [29]. In this work, we demonstrate validation of the PM from the CKM and create a refined model R-CKM using domain expert base validation criteria. As an integral part of HIS workflow (used as MLMs), R-CKM was validated against real patient test cases, and had an encouraging performance with an accuracy of 53.0%.

4.3. Validation of R-CKM: quality of MLMs

Using the PM as a primary source for the PM and evolving it into a refined model, R-CKM provides various advantages. First, R-CKM is created through a validation process from existing guidelines through CKM concepts. Second, R-CKM has the capability to directly integrate and work with healthcare workflows, which is not possible when guidelines are used as they exist. Third, R-CKM evolves from PM and has an accuracy of 53.0% on real patient data, which is encouraging, although it is relatively lesser than accuracy of the PM 59.0%. However it is expected that the accuracy of the R-CKM may be reduced due to evolution during refinements in the paths of PM. The accuracy of R-CKM is mainly dependent on the accuracy of PM and differences between the accuracies of PM and R-CKM dictate the gap between real practices and standard guideline-based practices. By investigating the detailed accuracy of each MLM from Fig. 8, it is revealed that candidate MLMs can be used in a real clinical setup. Currently, the MLMs are integrated with HIS, where resident doctors are testing 60 patient cases per week for further quality measurement of the generated recommendations. Although R-CKM is validated on a real dataset of 739 patients (offline), these MLMs will become part of the healthcare workflow after passing the (online) test phase. Despite the fact that the R-CKM model has many advantages, it cannot be considered complete because it works only for patient data readily available in HIS. Therefore, physicians may need to review the recommendations before selecting the final treatment plan.

4.4. Selection of CHAID over CRT

One of the main goals of this work was to convey the knowledge of how Arden Syntax can be used by novice physicians. Thus, the selection criteria was intended to build a PM that is simple enough with respect to involving minimal attributes and giving maximum accuracy. CHAID was ranked first even though CRT performed well on a given dataset because it gave fewer rules by involving fewer attributes. Selection of CHAID over CRT was only made to create simple R-CKM to be easily transformable to corresponding MLMs by physicians. However, CRT is considered a good candidate for R-CKM to work as a prediction machine, which involves other relevant attributes from the practices dataset.

4.5. R-CKM transformation into MLMs

The current approach selects the MLM-calling mechanism to transform the set of MLMs controlled via the root MLM. Importantly, the clinical logic can be distributed in various subMLMs depending on decision paths and choices of domain experts. Thus, R-CKM can be transformed into the different sets of MLMs mentioned in this work. For example, one option can be the merging of OralCavityComplexLv1 into RootMLM, whereby the other two subMLMs OralCavityComplexLv121 and OralCavityComplexLv122 are called by RootMLM.

4.6. vMR experiences in HL7 Arden Syntax

Senior and expert physicians are rarely attached to using advanced computer technology. Rather, they are concerned with work in routine clinics with minimal use of basic computer applications. In this study, we involved senior oncologists and young resident doctors. Resident doctors showed interest in the idea of executable knowledge creation, while senior oncologists were involved for validation of R-CKM. While interacting with resident doctors for transformation of MLMs, it was evident that they experienced difficulties with the comprehensive Arden Syntax specifications and were confused as to where to formally start writing MLMs. Knowledge engineers resolved this problem by creating short implementation guides, which covered the use of most important concepts for final MLMs. Second, physicians were interested in use of localized concepts in logic rather than digging into technical details of accessing vMR concepts. For example, (*secondLv1Treatment.procedureCode* = "413737006" AND *secondLv1Treatment.procedureMethod* != "387713003") should be written as *secondLv1TreatmentPlan* != "Surgery". This demand on the physician led back to the Arden Syntax curly brace problem; an approach that resolves the complexity of vMR is needed. For the current work, we introduced basic clinical statements, such as *Observation-Result*, *ProcedureEvent* and *Problem* to represent the local concepts used in R-CKM. To permanently address this problem, our ongoing research involves creating detailed clinical models of cancer from local practices and mapping the concepts into vMR concepts. The mapping will be plugged into a knowledge authoring tool as a bridge to provide abstractions for vMR concepts.

5. Conclusion and future work

The proposed work introduced a data-driven knowledge acquisition and validation approach using a three-phase process model whereby a R-CKM was created using a PM, which conforms to a CKM. The R-CKM is transformable to an executable knowledge base, where example executable MLMs are created for oral cavity cancer treatment plans. This approach of knowledge acquisition ensures that the created model is used for CDSS interventions with existing patient data schemas, which would otherwise not be feasible

with the direct implementation of general guidelines. The proposed approach encourages physicians to transform their professional experience into sharable clinical knowledge using HL7 Arden Syntax. Moreover, HL7 vMR was used in the creation of MLMs to avoid the intrinsic curly brace problem of HL7 Arden Syntax. Finally, the experience of domain experts with HL7 vMR was discussed and indicated the need for future work on standard simplification.

Future work will expand on the research presented here and will consider the knowledge verification techniques to verify the generated MLMs against R-CKM. Furthermore, completeness of R-CKM will be evaluated against the CKM with respect to available datasets and domain expert opinions.

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).

The authors would like to acknowledge the support of medical and IT staff of Shaukat Khanum Memorial Cancer Hospital and Research Center, Lahore especially; Dr. Raza Hussain, Dr. Sadaf, Dr. Hassan Iqbal, Mr. Idress Kookher, and Mr. Manzar Abbas.

Appendix A. Clinical treatment plan vMR detail clinical model

```
1 Package UCLab_SKM_OralCavityTreatmentPlan
2
3 imports VMR
4
5 class PatientClinicalStatementTreatments
6     treatmentIntent: ProcedureEvent
7     primaryLvlTreatment: ProcedureEvent
8     secondaryLvlTreatment: ProcedureEvent
9     tertiaryLvlTreatment: ProcedureEvent
10
11 class PatientClinicalStatementObservations
12     clinicalStageT: ObservationResult
13     clinicalStageN: ObservationResult
14     clinicalStageS: ObservationResult
15
16 class PatientClinicalStatementProblem
17     diseaseFrmHistofinding: Problem
18
19 class RecommendationRemarks
20     recommendedTreatmentCode: CS
21     remarks: ST
22
23 class Recommendation
24     treatment: ProcedureEvent
25     recommendationDetail: RecommendationRemarks
26
27 class RecommendationList
28     recommendationsList: set(Recommendation)
29
30 class CDSSInput
31     oralCavityTreatmentPlanEvent: CS
32     evaluatePatient: Patient
33     patientCompletedTreatments: PatientClinicalStatementTreatments
34     patientClinicalStaging: PatientClinicalStatementObservations
35     patientDisease: PatientClinicalStatementProblem
36
37 class CDSSOutput
38     oralCavityTreatmentPlanEvent: CS
39     evaluatePatient: Patient
40     recommendationsList: RecommendationList
41
42 EndPackage
```

Appendix B. Oral cavity: SNOMED codes for HIS concepts

Table B.8

HIS concept mapping with SNOMED concepts.

HIS concept	SNOMED concept	HIS designated values	SNOMED code (for values)	SNOMED description (for value)
Clinical stage T	385356007 Tumor stage finding (finding)	T0	58790005	T0 category (finding), T0 category, T0 stage, Tumor stage T0, Tumor stage T0
		T1	23351008	T1 category (finding), T1 category, T1 stage, Tumor stage T1, Tumor stage T1
		T2	67673008	T2 category (finding), T2 category, T2 stage, Tumor stage T2, Tumor stage T2
		T3	14410001	T3 category (finding), T3 category, T3 stage, Tumor stage T3, Tumor stage T3
		T4	65565005	T4 category (finding), T4 category, T4 stage, Tumor stage T4, Tumor stage T4
Clinical stage N	385382003 Node (category finding (finding), N stage finding, Node category finding, Node stage finding)	N0	62455006	N0 category (finding), N0 category, N0 stage, Node stage N0
		N1	53623008	N1 category (finding), N1 category, N1 stage, Node stage N1
		N2	46059003	N2 category (finding), N2 category, N2 stage, Node stage N2
		N3	5856006	N3 category (finding), N3category, N3 stage, Node stage N3
Clinical stage S	80631005 Clinical stage finding (finding), Clinical stage finding	I	13104003	Clinical stage finding, Tumor stage finding, Clinical stage I (finding)
		II	60333009	Clinical stage finding, Tumor stage finding, Clinical stage II (finding)
		III	50283003	Clinical stage finding, Tumor stage finding, Clinical stage III (finding)
		IV	2640006	Clinical stage finding, Tumor stage finding, Clinical stage IV (finding)
Treatment intent	395077000 Treatment intent (situation)	Radical	27762005	Radical procedure
		Palliative	363676003	Palliative – procedure intent
Histology description	250537006 Histopathology finding	Squamous cell carcinoma	402815007	Squamous cell carcinoma (disorder)
		Small cell carcinoma	74364000	Small cell carcinoma (morphologic abnormality)
		Carcinoma NOS	68453008	Carcinoma, no subtype (morphologic abnormality)
		Adenocarcinoma	35917007	Adenocarcinoma, no subtype (morphologic abnormality)
		Adenoid cystic carcinoma	11671000	Adenoid cystic carcinoma (morphologic abnormality)
		Adenoid cystic carcinoma	1338007	Basal cell carcinoma (morphologic abnormality)
		Squamous cell carcinoma in situ	59529006	Squamous cell carcinoma in situ, no International Classification of Diseases for Oncology subtype (morphologic abnormality)
		Verrucous carcinoma	89906000	Verrucous carcinoma (morphologic abnormality)
		Malignant melanoma	2092003	Malignant melanoma, no International Classification of Diseases for Oncology subtype (morphologic abnormality)
		Pleomorphic adenoma	8360001	Pleomorphic adenoma (morphologic abnormality)
		Spindle cell carcinoma	65692009	Spindle cell carcinoma (morphologic abnormality)
		Ameloblastoma, malignant	88253001	Ameloblastoma, malignant (morphologic abnormality)
		Adenoid squamous cell carcinoma	85956000	Adenoid squamous cell carcinoma (morphologic abnormality)
		nasopharyngeal carcinoma	449248000	Nasopharyngeal carcinoma (disorder)
		Sebaceous adenocarcinoma	54734006	Sebaceous adenocarcinoma (morphologic abnormality)
Treatment plan	1.413737006 Cancer hospital treatment completed (situation) 2.225292002 Developing a treatment plan (procedure)	Sarcoma, not otherwise specified	397355008	Dendritic cell sarcoma, not otherwise specified (morphologic abnormality)
		Plasmacytoma, not otherwise specified	415112005	Plasmacytoma (disorder)
		Mucoepidermoid carcinoma	4079000	Mucoepidermoid carcinoma (morphologic abnormality)
		Chemotherapy	367336001	Chemotherapy (procedure)
		CRT (Chemoradiotherapy)	703423002	Combined chemotherapy and radiation therapy (procedure)
		RT (Radiotherapy)	108290001	Radiation oncology AND/OR radiotherapy (procedure)
		Surgery	387713003	Surgical procedure (procedure)
		Induction Chemotherapy	450827009	Induction chemotherapy (procedure)

Appendix C. Arden Syntax implementation guide

Following are details of selected HL7 Arden Syntax artifacts used in this study.

- *object*: It is an important construct for representing clinical data in vMR object format. Physicians are introduced to methods of using *new.. with* for object creation.
- *EXTRACT ATTRIBUTE NAMES*: It is used to retrieve detailed information of data members of vMR objects.
- *ATTRIBUTE FROM*: The Attribute From operator is used to access data members of vMR objects.
- *List*: It is used to access data members of complex vMR objects and multiple recommendations are stored in the list. Furthermore, a

Merge operator is used to merge multiple recommendations into a common list.

- *Control Structures*: *IF*, *IF-ELSE* and *For-Do* control structure are used for creation of logic and iterating a recommendation list.
- *Call With*: The design of MLMs include subMLMs, which are called by other MLMs. Thus, the *Call With* operator is introduced for invocation of MLMs. Furthermore, the *ARGUMENT* concept is introduced to handle the data among calling mechanism of MLMs.
- *Logical operators*: This set of logical operators is introduced and includes *OR*, *AND*, *IN* and *IS [not]*.
- *Select From Where*: For data importing from HIS, the *select from where* operator was used that populate the vMR based clinical statements.

Appendix D. Oral cavity primitive root MLM

```

1 maintenance:
2 title: Localized Treatment Plan for Oral Cavity Cancer: Primitive MLM;;
3 mlname: OralCavity_LocalizedPrimitive;;
4 arden: ASTM-E1460-1995;;
5 version: 2.7;;
6 institution: UC Lab and SKMCH;;
7 author: Mr. Maqbool Hussain;;
8 specialist: Dr. Muhammad Irfan and Dr. Hassan Iqbal ;;
9 date: 2014-05-30;;
10 validation: Prototype Testing;;
11 library:
12 purpose: OralCavity_LocalizedPrimitive will trigger for treatment plan of
           oral cavity cancer as prototype in local clinical set-up for head and
           neck cancer;;
13 explanation: This MLM is evoked when patient is provided in encounter for
           proposition of treatment plan having oral cavity cancer. The result of
           this MLM is recommendation of treatment plan based on patient ;;
14 keywords: oral cavity cancer; treatment plan; NCCN Guidelines;;
15 citations: NCCN Guidelines and SKMCH treatment knowledge model from data;;

```

```
16
17 knowledge:
18   type: data-driven;;
19
20 data:
21   oralCavityTreatment_plan := event {treatment_plan where class =
      oralCavityTreatmentPlanEvent};
22   PatientClinicalStatementTreatments := object [ProcedureEvent,
      ProcedureEvent, ProcedureEvent, ProcedureEvent];
23   PatientClinicalStatementObservations := object [ObservationResult,
      ObservationResult, ObservationResult];
24   PatientClinicalStatementProblem := object [Problem];
25   RecommendationRemarks := object [CS,ST];
26
27   patientCompletedTreatments := read as PatientClinicalStatementTreatments
28   { select ProcedureEvent from client where ProcedureEvent.procedureCode = "
      413737006" OR ProcedureEvent.procedureCode = "395077000" };
29
30   patientClinicalStaging := read as PatientClinicalStatementObservations
31   { select ObservationResult from client where ObservationResult.
      observationFocus IN ( "385356007", "385382003", "80631005" ) };
32
33   patientDisease := read as PatientClinicalStatementProblem
34   { select Problem from client where Problem.problemCode = "250537006" };
35
36   /* Treatment Recommendations as result of patient status */
37   Recommendation := object[ ProcedureEvent, RecommendationRemarks ];
38   recommendationList := (); /* List of Recommendations */
39
40   /*.....Child MLM declarations .....*/
41   mlmOralCavityComplexLv11:= MLM 'mlmOralCavityComplexLv11.mlm'
42   ;;
43
44 evoke:
45   oralCavityTreatment_plan;;
46
47 logic:
48
49   /* Using "Extract Attribute Names" operator. It returns all names of
      attributes as string list for given object */
50   /* patClinicalStagingList := EXTRACT ATTRIBUTE NAMES patientClinicalStaging
      ; */
51   patCompletedTreatmentsList := EXTRACT ATTRIBUTE NAMES
      patientCompletedTreatments;
52
53
54   /* Retrieving Treatment Intent */
55   treatmentIntent := ATTRIBUTE patCompletedTreatmentsList[1] FROM
      patientCompletedTreatments;
56   firstLv1Treatment := ATTRIBUTE patCompletedTreatmentsList[2] FROM
      patientCompletedTreatments;
57
58   IF (treatmentIntent.procedureCode = "395077000" AND treatmentIntent.
      procedureMethod = "363676003") THEN
59     IF ( (firstLv1Treatment IS NULL) OR (firstLv1Treatment.procedureCode = "
      413737006" AND treatmentIntent.procedureMethod != "108290001") ) THEN
60       plannedTreatment := new ProcedureEvent with "225292002" , "108290001";
```

```
61  plannedTreatmentRemarks := new RecommendationRemarks with "225292002", "
    Radiotherapy is recommended treatment plan !!!";
62  treatmentRecommendation := new Recommendation with plannedTreatment,
    plannedTreatmentRemarks;
63  recommendationList := recommendationList, treatmentRecommendation;
64
65  ELSE
66
67  plannedTreatmentRemarks := new RecommendationRemarks with "413737006", "
    Radiotherapy is already done, evaluate patient case for further
    treatment !!!";
68  treatmentRecommendation := new Recommendation with firstLvlTreatment,
    plannedTreatmentRemarks;
69  recommendationList := recommendationList, treatmentRecommendation;
70
71  END IF;
72
73  CONCLUDE TRUE;
74
75  ELSEIF (treatmentIntent.procedureCode = "395077000" AND treatmentIntent.
    procedureMethod = "27762005") THEN
76  IF ( (firstLvlTreatment IS NULL) OR (firstLvlTreatment.procedureCode = "
    413737006" AND treatmentIntent.procedureMethod != "450827009") ) THEN
77
78  plannedTreatment := new ProcedureEvent with "225292002", "450827009";
79  plannedTreatmentRemarks := new RecommendationRemarks with "225292002", "
    Induction Chemotherapy is recommended treatment plan !!!";
80  treatmentRecommendation := new Recommendation with plannedTreatment,
    plannedTreatmentRemarks;
81  recommendationList := recommendationList, treatmentRecommendation;
82
83  ELSE
84
85  plannedTreatmentRemarks := new RecommendationRemarks with "413737006", "
    Induction Chemotherapy is already done !!!";
86  treatmentRecommendation := new Recommendation with firstLvlTreatment,
    plannedTreatmentRemarks;
87  recommendationList := recommendationList, treatmentRecommendation;
88
89  /*CALLING SUB MLM: */
90  recommendationLvl1List := call mlmOralCavityComplexLvl1 with
    patientCompletedTreatments, patientClinicalStaging, patientDisease;
91  recommendationList := recommendationList MERGE recommendationLvl1List;
92
93  END IF;
94
95  CONCLUDE TRUE;
96
97  ELSE
98
99  CONCLUDE FALSE;
100
101  END IF;;
102
103  action:
104
105  FOR recommendations IN recommendationList DO
106
107  WRITE recommendations.ProcedureEvent.procedureMethod || recommendations.
    RecommendationRemarks;
108
109  END DO;
110  ;;
111
112  End;;
```

References

- [1] Wright A, Sittig DF. A framework and model for evaluating clinical decision support architectures. *J Biomed Inf* 2008;41(6):982–90.
- [2] Wright A, Sittig DF, Ash JS, Sharma S, Pang JE, Middleton B. Clinical decision support capabilities of commercially-available clinical information systems. *J Am Med Inf Assoc* 2009;16(5):637–44.
- [3] National comprehensive cancer network; 2014 <http://www.nccn.org/> [accessed 24.04.15].
- [4] Samwald M, Fehre K, De Bruin J, Adlassnig K-P. The Arden Syntax standard for clinical decision support: experiences and directions. *J Biomed Inf* 2012;45(4):711–8.
- [5] HL7CDS.WG, HL7 version 3 standard: virtual medical record for clinical decision support (vMR-CDS) logical model. Ann Arbor, MI: Health Level Seven, Inc.; 2014 (Release 2).
- [6] IHTSDO. Snomed ct; 2014 <http://www.ihtsdo.org/snomed-ct/> [accessed 26.12.14].
- [7] Peleg M. Computer-interpretable clinical guidelines: a methodological review. *J Biomed Inf* 2013;46(4):744–63.
- [8] Peleg M, Gutnik LA, Snow V, Patel VL. Interpreting procedures from descriptive guidelines. *J Biomed Inf* 2006;39(2):184–95.
- [9] Serban R, Ten Teije A. Exploiting thesauri knowledge in medical guideline formalization. *Methods Inf Med* 2009;48(5):468–74.
- [10] Hripcsak G, Cimino JJ, Johnson SB, Clayton PD. The columbia-presbyterian medical center decision-support system as a model for implementing the Arden Syntax. In: *Proceedings of the annual symposium on computer application in medical care*. 1991. p. 248–52.
- [11] Perera S, Henson C, Thirunarayan K, Sheth A, Nair S. Semantics driven approach for knowledge acquisition from EMRs. *IEEE J Biomed Health Inf* 2014;18(2):515–24.
- [12] Gomo V, Vida M, Robu R, Stoicu-Tivadar V, Bernad E, Lupse O. From data mining rules to medical logical modules and medical advices. *Stud Health Technol Inf* 2012;192:1094.
- [13] Toussi M, Lamy J-B, Le Toumelin P, Venot A. Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. *BMC Med Inf Decis Mak* 2009;9(1):28.
- [14] Shalom E, Shahar Y, Taieb-Maimon M, Bar G, Yarkoni A, Young O, et al. A quantitative assessment of a methodology for collaborative specification and evaluation of clinical guidelines. *J Biomed Inf* 2008;41(6):889–903.
- [15] Miller PL. Domain-constrained generation of clinical condition sets to help test computer-based clinical guidelines. *J Am Med Inf Assoc* 2001;8(2):131–45.
- [16] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inf* 2008;77(2):81–97.
- [17] Pombo N, Araújo P, Viana J. Knowledge discovery in clinical decision support systems for pain management: a systematic review. *Artif Intell Med* 2014;60(1):1–11.
- [18] Lorena AC, Jacintho LF, Siqueira MF, Giovanni RD, Lohmann LG, De Carvalho AC, et al. Comparing machine learning classifiers in potential distribution modelling. *Expert Syst Appl* 2011;38(5):5268–75.
- [19] IBMSPSS.User.Guide. IBM SPSS decision trees 21; 2012 <http://www.sussex.ac.uk/its/pdfs/SPSS.Decision.Trees.21.pdf> (accessed 24.12.14).
- [20] de Clercq PA, Blom JA, Korsten HH, Hasman A. Approaches for creating computer-interpretable guidelines that facilitate decision support. *Artif Intell Med* 2004;31(1):1–27.
- [21] ArdenSyntax.WG. Health level seven Arden Syntax for medical logic systems. Ann Arbor, MI: Health Level Seven, Inc.; 2014 (Version 2.10).
- [22] Kawamoto K, Del Fiol G, Strasberg HR, Hulse N, Curtis C, Cimino JJ, et al. Multi-national, multi-institutional analysis of clinical decision support data needs to inform development of the hl7 virtual medical record standard. In: *AMIA annual symposium proceedings*, vol. 2010. 2010. p. 377–81.
- [23] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. *ACM SIGKDD Explor Newsl* 2009;11(1):10–8.
- [24] Hussain M, Afzal M, Khan WA, Lee S. Clinical decision support service for elderly people in smart home environment. In: *2012 12th International Conference on control automation robotics & vision (ICARCV)*, IEEE. 2012. p. 678–83.
- [25] Hussain M, Ali T, Khan WA, Afzal M, Lee S, Latif K. Recommendations service for chronic disease patient in multimodel sensors home environment. *Telemed e-Health* 2015;21(3):185–99.
- [26] Akbar A, Bhatti A, Syed A, Khattak S, Jamshed A, Kazmi S. P10 induction chemotherapy followed by chemoradiotherapy before surgery for rectal adenocarcinoma. *Eur J Cancer* 2014;50(1):S12.
- [27] Bottrighi A, Giordano L, Molino G, Montani S, Terenziani P, Torchio M. Adopting model checking techniques for clinical guidelines verification. *Artif Intell Med* 2010;48(1):1–19.
- [28] Pérez B, Porres I. Authoring and verification of clinical guidelines: a model driven approach. *J Biomed Inf* 2010;43(4):520–36.
- [29] Ten Teije A, Marcos M, Balser M, van Croonenborg J, Duelli C, van Harmelen F, et al. Improving medical protocols by formal methods. *Artif Intell Med* 2006;36(3):193–209.