Original Research

Smart Extraction and Analysis System for Clinical Research

Muhammad Afzal, MS,[1] Maqbool Hussain, PhD,[1] Wajahat Ali Khan, PhD,[1] Taqdir Ali, MS,[1] Arif Jamshed, MBBS, FRCR,[2] and Sungyoung Lee, PhD[1]

[1]Ubiquitous Computing Lab, Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea.
[2]Shaukat Khanum Memorial Cancer Hospital and Research Center, Lahore, Pakistan.

Abstract

Background: *With the increasing use of electronic health records (EHRs), there is a growing need to expand the utilization of EHR data to support clinical research. The key challenge in achieving this goal is the unavailability of smart systems and methods to overcome the issue of data preparation, structuring, and sharing for smooth clinical research.* ***Materials and Methods:*** *We developed a robust analysis system called the smart extraction and analysis system (SEAS) that consists of two subsystems: (1) the information extraction system (IES), for extracting information from clinical documents, and (2) the survival analysis system (SAS), for a descriptive and predictive analysis to compile the survival statistics and predict the future chance of survivability. The IES subsystem is based on a novel permutation-based pattern recognition method that extracts information from unstructured clinical documents. Similarly, the SAS subsystem is based on a classification and regression tree (CART)-based prediction model for survival analysis.* ***Results:*** *SEAS is evaluated and validated on a real-world case study of head and neck cancer. The overall information extraction accuracy of the system for semistructured text is recorded at 99%, while that for unstructured text is 97%. Furthermore, the automated, unstructured information extraction has reduced the average time spent on manual data entry by 75%, without compromising the accuracy of the system. Moreover, around 88% of patients are found in a terminal or dead state for the highest clinical stage of disease (level IV). Similarly, there is an ~36% probability of a patient being alive if at least one of the lifestyle risk factors was positive.* ***Conclusion:*** *We presented our work on the development of SEAS to replace costly and time-consuming manual methods with smart automatic extraction of information and survival prediction methods. SEAS has reduced the time and energy of human resources spent unnecessarily on manual tasks.*

Keywords: *information extraction, e-health, pattern recognition, clinical research, cancer survival analysis*

Introduction

Cancer is a major public health problem worldwide,[1] making it the second leading cause of death in the United States, resulting in one in four deaths.[2] A recent review study indicated that one in three people in the United Kingdom develop some form of cancer during their lifetime.[3] Head and neck cancer (HNC) represents a large, heterogeneous group with ~460,000 cases worldwide,[4] and it is often treated by an intensive combination of surgery, radiotherapy, and chemotherapy.[5] In the last 20 years, even though innovative methods have been developed for early detection and treatment, which helped decrease the cancer-related death rate, cancer is still a major cause of concern in the United States.[6] With the increased global use of electronic health records (EHRs), there is a growing need to expand the utilization of EHR data to support clinical research.[7] A major challenge faced during this process is the transformation of clinical narratives to a structured format to allow development of innovative analysis services of the data.

In the biomedical domain, much of the available clinical data are recorded as freestyle text in the form of clinical documents.[8] Approximately 96% of cancer diagnoses originate in a surgical pathology laboratory,[3] and this can be considered an important source of information to help in the treatment of patients with cancer. This free text is convenient for describing clinical activities, but it is hard to use for searching, statistical analysis, or decision support. To improve the overall quality of care by using health analytics features, we need to design a bridge solution for automating the process of accessing data from various data sources. The bridge solution narrows the knowledge utilization gap between the source and target systems, through which the source system creates knowledge in the form of unstructured documents, and the target system utilizes this knowledge for different purposes. Several different techniques are used to extract information based on symbolic information, statistical methods, or machine learning. In the clinical domain, information extraction helps clinicians to answer questions, such as *How many patients in each clinical or pathological stage do we serve? How many patients were diagnosed with primary cancer*

in year x? What percentage of these patients had metastatic tumors in the head? The studies and reviews on text mining[9] and on information extraction[10] have discussed the existing tools and techniques of information extraction in the biomedical and clinical domains. A recent review discussed the current status and future directions of text mining in the area of cancer-related information.[3]

Various automated and semiautomated methods have been designed for knowledge acquisition from clinical documents.[11] Similarly, different dictionary- and rule-based techniques have been designed to extract information from electronic medical records (EMRs).[12] A semiautomatic data extraction approach is used to obtain information on the quality of prescribed medication in general practice, in a setting for different EMR-software systems.[13] An automated medication extraction system (MedEx)[14] accurately extracted medication names and signatures from clinical narratives, with F-measures greater than 90% on a set of 25 clinic visit notes. Campbell and Johnson wrote in favor of dependency grammar for biomedical text because of the lack of sentence grammar.[15] Spell checking is a useful task during preprocessing as physicians often misspell medical and clinical terms when writing clinical notes. The misspelling levels in medical records are about 10% higher than the misspelling levels for other types of texts.[16]

On the contrary, predictive data mining in clinical medicine is summarized by reviewing the current issues and guidelines.[17] The author of that study mentioned that the data drawn from heterogeneous sources are required to be integrated in the construction of reliable predictive models. Delen et al. used two popular data mining algorithms (artificial neural networks and decision trees) to develop predictive models for breast cancer survivability.[6] They uploaded the raw data into an MS Access database, an SPSS statistical analysis tool, a statistical data miner, and the Clementine data mining toolkit.[6] Recently, the focus has changed to mining data from EHRs, which have the potential to establish a new set of patient-stratification principles and to reveal unknown disease correlations.[18] The author of that study mentioned that the text in clinical narratives is considered to be a cornerstone for ensuring informed decision-making.

In this article, we present the proposed smart extraction and analysis system (SEAS) that consists of two subsystems: (1) the information extraction system (IES) and (2) the survival analysis system (SAS). The key concept behind the IES is based on an innovative method of permutation-based pattern recognition (PR) that extracts information from unstructured clinical documents. In the same way, SAS is used for a descriptive and predictive analysis (PA) to derive descriptive statistics of the death ratio and provide grounds for predicting the future chance of patient survivability. The system is based on classification and regression tree (CART)-based data analysis and an analytics generation model.

The proposed SEAS was implemented in a research centre of Shaukat Khanum Memorial Cancer Hospital & Research Center (SKMCH&RC*) that performs cancer management, an in-house hospital information system was previously developed to automate patient record management.[19] We sought close support from domain experts in information extraction for manual annotation and verification of the results modeled through extraction and prediction methods for use in clinical research. The main contributions of this work are divided into the following areas: (1) design and development of an IES that supports an innovative method of permutation-based PR from unstructured documents and (2) development of a CART-based SAS for the identification and predication of patient survivability.

Background and Motivation

Clinical data are increasing exponentially, which requires smart technology to deliver services based on the data. One of the bottlenecks in development of such services is the variety of formats of the data. Processing these unstructured data formats can be a very expensive procedure. It is therefore necessary to have all of the data in a structured format so that the analytics service, decision support service, and many other services can be provided to the stakeholders. We developed the SEAS to collect EHR data from a hospital information system and convert it to a structured format, which is utilized for survival analysis. Collectively, this allows a cost-effective environment suitable for clinical and epidemiological research.

Our motivation for the proposed work can be categorized by three unavoidable factors in this technologically advanced world: time, cost, and services. Considering the time factor, we use the example of resident doctors manually converting clinical notes from an unstructured format to a structured

*Shaukat Khanum Memorial Cancer Hospital & Research Centre (SKMCH&RC), Lahore, Pakistan, is a dynamic, state of the art cancer hospital providing comprehensive care free of cost to thousands of inhabitants of Pakistan. The hospital provides excellent services with support from donations of well wishers throughout the world. The hospital has developed an indigenous comprehensive Hospital Management Information System (HMIS) that includes all of the diverse workflows of the hospitals. The HMIS provides automation for factors from patient encounters to laboratory results and pharmacy (www.shaukatkhanum.org.pk/home.html).

format. A considerable amount of time and energy can be saved, and the SEAS can benefit other areas by automatically converting the semistructured and unstructured data formats into a structured format. This will only require doctors to verify that the conversion is correct. Another aspect is related to the cost of managing the conversion. The proposed system can perform the tasks of many resources involved in the conversion process. This job is now performed with a single click, which reduces the expenditure of many resources. Finally, the conversion and storage of unstructured data into structured data are largely helpful for the different services provided to the different stakeholders. Researchers can directly utilize the structured data to generate various kinds of analytics-related services on the data. Physicians, nurses, or other clinical researchers can benefit from these services in their research and decision-making process.

Materials and Methods

We implemented the SEAS in a research center that is connected to a central unit that utilizes a comprehensive hospital management information system (HMIS) for the management, diagnosis, and treatment of different types of cancer. The HMIS provides services to different departments such as pathology, radiology, and surgery to achieve coordinated care of diseases. As highlighted in *Figure 1*, the central unit is connected to the remote research center and to other remote units, including a diagnostic center, collection point, and walk-in clinic. The research center, which is the main focus of this study, accesses EHR data from the HMIS and provides analysis services on different specialties, such as HNC and breast cancer.

PROPOSED SYSTEM ARCHITECTURE

The SEAS consists of two subsystems: the *IES* and the *SAS*. As shown in *Figure 2*, the IES has a data handler module that collects EHR data from the central unit. These EHR data are composed of three types: structured, semistructured, and unstructured. The semistructured and unstructured parts of the data are passed to the *natural language preprocessing* (*NLPreP*) module to perform the basic NL functions such as section identification, stop word removal, and normalization. The preprocessed data are sent to the *named entity recognition* (*NER*) module to identify the concepts of interest with the help of a domain *lexicon*. For the unstructured part of the data, mainly the *PR* is required to identify the values to assign to the correct attributes. Once semistructured or unstructured data are properly transformed to a structured form, they are stored in the *intermediate database* (*IDB*).

The structured data in the IDB are accessible to SAS, which implements a query manager (QM) that queries the IDB to fetch data according to the requirements.

The descriptive analysis (DA) module provides statistics on patient data to identify the cancer incidences and death ratio. The PA module provides the classification of data and predicts

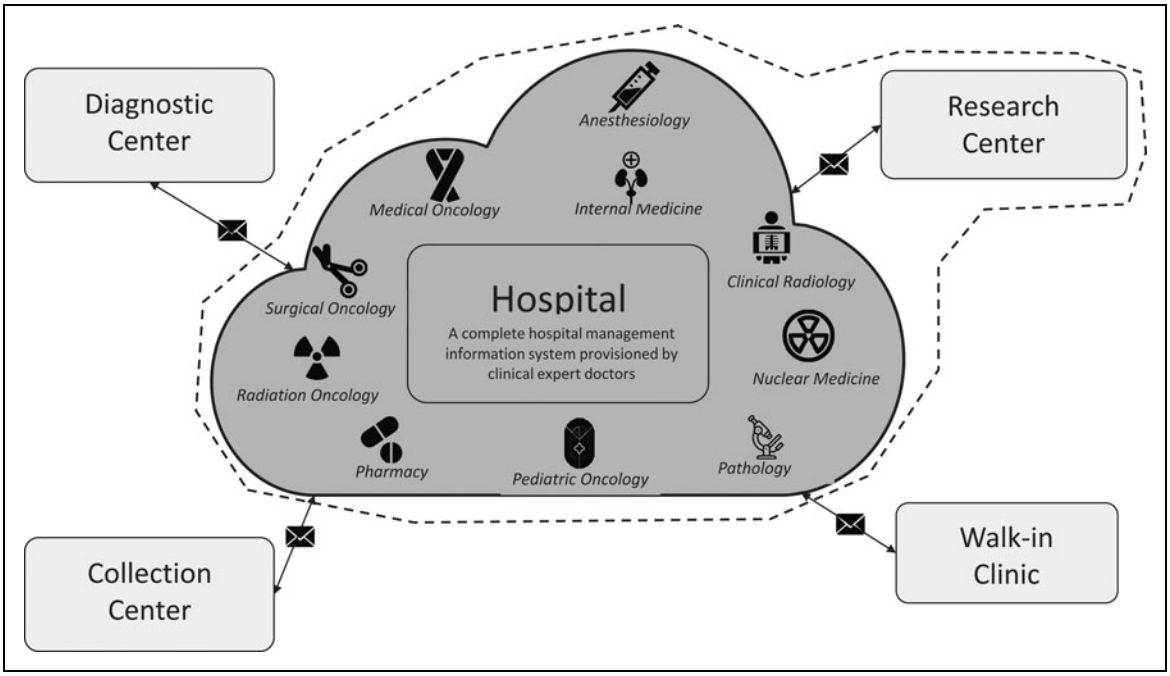

Fig. 1. Overview of communication infrastructure of the smart extraction and analysis system.

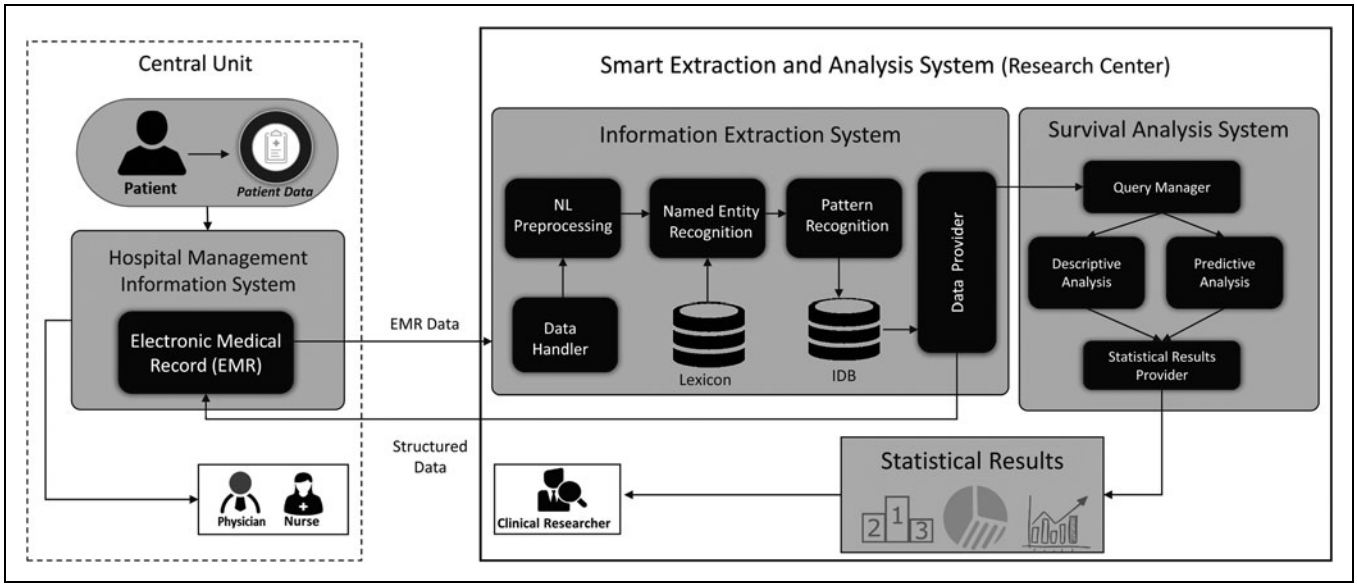

Fig. 2. Functional architecture of the smart extraction and analysis system connected to the hospital management information system (central unit).

the chance of survivability of future patients. The statistical results generation (SRG) module provides different filters to include/exclude data of the researcher's choice. The final statistical results are available for a clinical researcher to investigate the insights of patient data.

COMMUNICATION PATTERN

The interaction of SEAS with the HMIS in the central unit is performed through a Web service that retrieves the requested documents. The document retrieval function acquires the data from the source system through the RESTful Web service using the JSON format.[20] The central unit implements the EMR as a part of the HMIS. The HMIS collects heterogeneous types of data, including patient data. These patient data are sent to SEAS, which processes and structures them using IES and stores the data in the IDB. The structured data in the IDB are accessible to SAS and is also communicated to the HMIS of the central unit for analysis.

INFORMATION EXTRACTION SYSTEM

The IES is a modular system that uses three main components, NLPreP, NER, and PR, which are supported by three other components: the data handler, lexicon, and IDB. IES receives EMR data from the central unit through its data handler component. The EMR data consist of three parts: structured, semistructured, and unstructured. Each of these parts is processed according to its format. As described in *Figure 3*, structured data are directly passed to the IDB without any further information extraction. Semistructured data are passed through the data handler, NLPreP, and NER. PR is required for unstructured data in addition to the steps of semistructured data, because of the complexity of different data

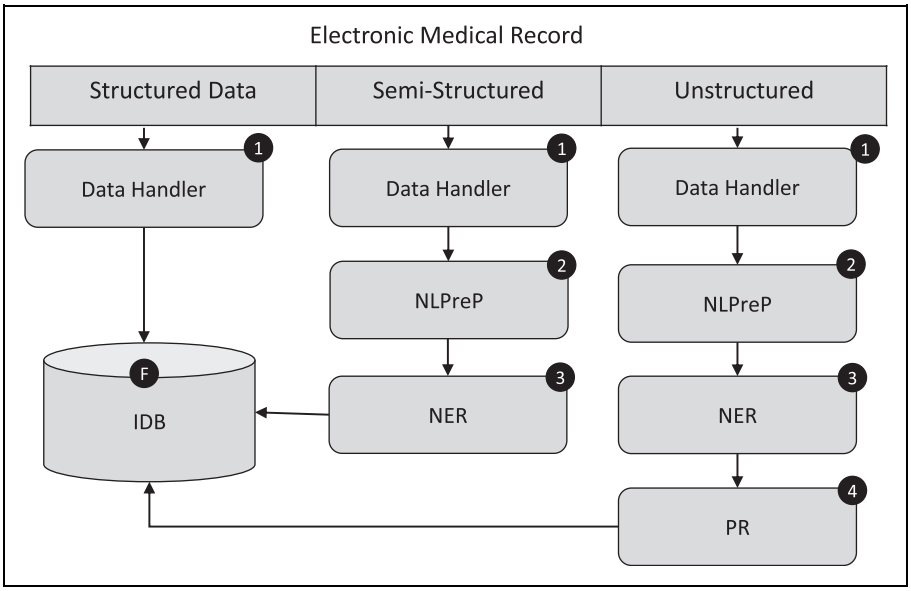

Fig. 3. Flow diagram of electronic health record data processing (F = final). IDB, intermediate database; NER, named entity recognition; NLPreP, natural language preprocessing; PR, pattern recognition.

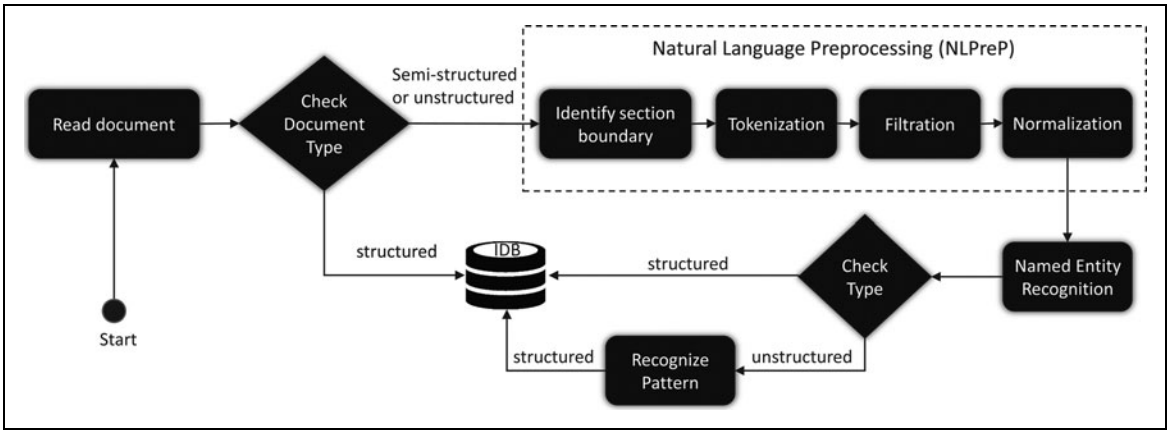

Fig. 4. Natural language preprocessing steps.

formats. We present the methods implemented for NLPreP, NER, and PR in the subsequent sections.

Natural language preprocessing. Four methods are involved in the preprocessing. First, the boundary of the section is identified to identify the start and endpoints of the text. In the second step, the text is tokenized; this is followed by the removal of unnecessary information in the third step. Finally, in the fourth step, text tokens are normalized for easy identification of the named entities. The preprocessing steps are described in the highlighted portion of *Figure 4*. After assessing the type of inputted document, if it is either semistructured or unstructured, the NLPreP module identifies the section on the basis of predefined keywords, which are used as start and end markers, similar to the data-driven approach of Denny et al.[21,22] The identified section is tokenized to remove the stop words, special characters (except the space character), and symbols such as brackets, curly brackets, parentheses, commas, and sentence terminators. The remaining tokens are normalized by regulating the alphabetic case, spelling variations, and mentions.

Named entity recognition. NER identifies the concepts of interest, which are domain specific and require a domain lexicon. We developed a lexicon for the concepts used in HNC histology reports. The lexicon for mapping lymphatic information is partially shown in *Table 1*. There are two kinds of concepts stored in the dictionary: *concepts with values* and *concepts without values*. *Concepts with values* refer to those for which values are meaningful, such as Lymph Node Level 1 having a value of x. *Concepts without values* refer to those concepts for which values are meaningless, such as positive or negative. Such concepts in the lexicon are identified with a flag value attribute called "ValueYNID." A flag value of 1 is used for *concepts with values*. Similarly, multiple words in a concept are identified with "MultipleWords" in the lexicon.

Table 1. Partial Lymph Node Mapping Lexicon with Concepts and Meta-Information

CONCEPTDESC	LEVELTYPE	REGULAREXPRESS...	VALUEYNID	MULTIPLEWORDS
Level 1	1	^[0-9]+$	1	2
Level 2	2	^[0-9]+$	1	2
Level 3	3	^[0-9]+$	1	2
Level 4	4	^[0-9]+$	1	2
Level 5	5	^[0-9]+$	1	2
Level-1	1	^[0-9]+$	1	*NULL*
Level-2	2	^[0-9]+$	1	*NULL*
Level-3	3	^[0-9]+$	1	*NULL*
Level-4	4	^[0-9]+$	1	*NULL*
Level-5	5	^[0-9]+$	1	*NULL*
Level-I	1	^[0-9]+$	1	*NULL*
Level-II	2	^[0-9]+$	1	*NULL*
Level-III	3	^[0-9]+$	1	*NULL*
Level-IV	4	^[0-9]+$	1	*NULL*
Level-V	5	^[0-9]+$	1	*NULL*

In the course of lexicon matching, two types of situations arise: an exact match and a partial match. In an exact match, a single word in the lexicon exactly matches one of the extracted words in the text. In a partial match, if a word in the lexicon matches two or more words in the list of recognized meaningful words, the two extracted words are concatenated to produce a single word and then that word is compared to the lexicon. However, this rule applies only when the concepts occur adjacent to each other.

Permutation-based PR. NER identifies meaningful words in the text; however, situations occur where entity identification cannot determine which value should be assigned to which attribute. The assignment of values to the corresponding attributes often follows a specific template. The complex part of populating a template is the correct identification of corelated information that occurs in a section, more specifically, assigning correct values to the attributes.

In text, some concepts appear with non-numeric values and some appear with numeric values. The concepts that appear with numeric values are known as key-value concepts. With the lymphatic system, all of the levels are considered keys because all of them have numeric values; thus, all are written in the key-value format. The sequence of concepts plays an important role in the correct assignment of values to the keys. In some cases, a value follows its key, while in other cases, a key follows its value. Sometimes, all keys appear first and are followed by their values, and sometimes all values appear first, followed by their keys. In the lymphatic system, since the values for different levels are numeric, the keys and their values occur in a specialized form. We approached this problem of discovering the pattern of keys and the occurrence of their values using the permutation method described in Eq. (1).

Let n represent the size of the set of keys from which permutations are derived and r represent the size of each permutation. The permutation is the arrangement of keys that occurred in the text.

$$\sum_{i=1}^{k} (P(n, r_i) = \sum_{i=1}^{k} \left(\frac{n!}{(n-r_i)!} \right), \quad (1)$$

For $n = 4$, that is, two keys with two corresponding values, we obtain the following results:

$$(P(n, r_i) = P(4,4) + P(4,3) + P(4,2) + P(4,1) = 64.$$

Not all of the 64 permutations are required to extract information from the narratives. Based on observations and heuristics, these 64 permutations create certain patterns, which can be either a paired key-value set or an unpaired key-value set. The paired key-value set contains patterns that have an equal number of keys and values. The unpaired key-value set contains patterns that have unpaired keys and values. In both paired and unpaired patterns, keys are distinguishable from one another. We developed the key-value assignment algorithm, Algorithm 1, to associate values with their correct keys.

Algorithm 1. Key-Value Assignment Algorithm

```
Input:
    keyValueSequence; //input string consists of keys and values
Output:
    assignedKeyValues; //resolved key value sequence
for all sequences in keyValueSequence
    keys = { K1, K2, K3, ..., Kn} ← findKeys(sequence);
    values = { V1, V2, V3, ..., Vn} ← findKeys(sequence);
    keysCount ← countKeys(keys);
    valuesCount ← countValues(values);
    permutationList ← generatePermutation(keysCount, valuesCount);
    // using Eq. (1).
    permutation ← MatchPermutation(sequence, permutationList);
    pattern = FindPattern(permutation);
    if (keysCount = valuesCount)
        assignedKeyValues ← resolvePattern(pattern, Rules);
        //rule 1–4 in Table 2
    else if (keysCount != valuesCount)
        assignedKeyValues ← resolvePattern(pattern, Rules); //rule 5–8
        in Table 2
    return assignedKeyValues;
end for
```

The algorithm searches for permutations. If a permutation is found, it counts the number of keys and values. If the numbers are equal, it looks up the corresponding rule in *Table 2*. For instance, rule 1 is used to resolve the permutation. It starts with the first key on the left and assigns the first value to it. It continues this process until it reaches the last value.

Storage of structured information. The extracted information from structured, semistructured, and unstructured documents is collectively stored in the IDB in a structured format. A representation of the IDB in the form of a relational database is shown in *Figure 5*.

For efficient retrieval, the design of the IDB is made in such a way as to divide all information in logical tables, for example,

Table 2. Pattern Parsing Rule Descriptions

RULE	NAME	DESCRIPTION
Rule 1	Key-value exact	The value is assigned to the key appearing in the first part.
Rule 2	Value-key exact	The value is assigned to the key appearing in the following position.
Rule 3	All keys followed by all values	The key appearing in the first position is assigned the value appearing in the first position, the second key is assigned the second value, and so on.
Rule 4	All values followed by all keys	The key appearing in the first position is assigned the value appearing in the first position, the second key is assigned the second value, and so on.
Rule 5	Key-value nearest neighbor	The value is assigned to the key that is nearest to it.
Rule 6	Value-key nearest neighbor	The value is assigned to the nearest key in the following position.
Rule 7	All keys	Unresolved pattern
Rule 8	All values	Unresolved pattern

"PatientDiagnosis," "PatientSurgery," "PatientRadiotherapy," and "PatientStage," connected to a main table "PatientEncounters," which contains the administrative attributes such as the date of an encounter and identifiers. This kind of structure avoids the second-level joining that can create delay in the saving and retrieval processes.

SURVIVAL ANALYSIS SYSTEM

The structured information in the IDB is accessible to SAS and is used to derive different types of analytics to gain insight into the data and their relationships, which is helpful for analyzing individual as well as population behavior. SAS is composed of four modules: QM, DA, PA, and SRG. QM and SRG are types of communication modules. The QM performs a transformation function that transforms the data based on the user-provided query in an SPSS[23]-compliant format. SRG compiles and generates the statistical results and provides them to the clinical researchers. Technically, DA and PA comprise the core part of SAS; these are discussed in the subsequent subsections. Using the SAS system, two types of analysis: descriptive and predictive, are

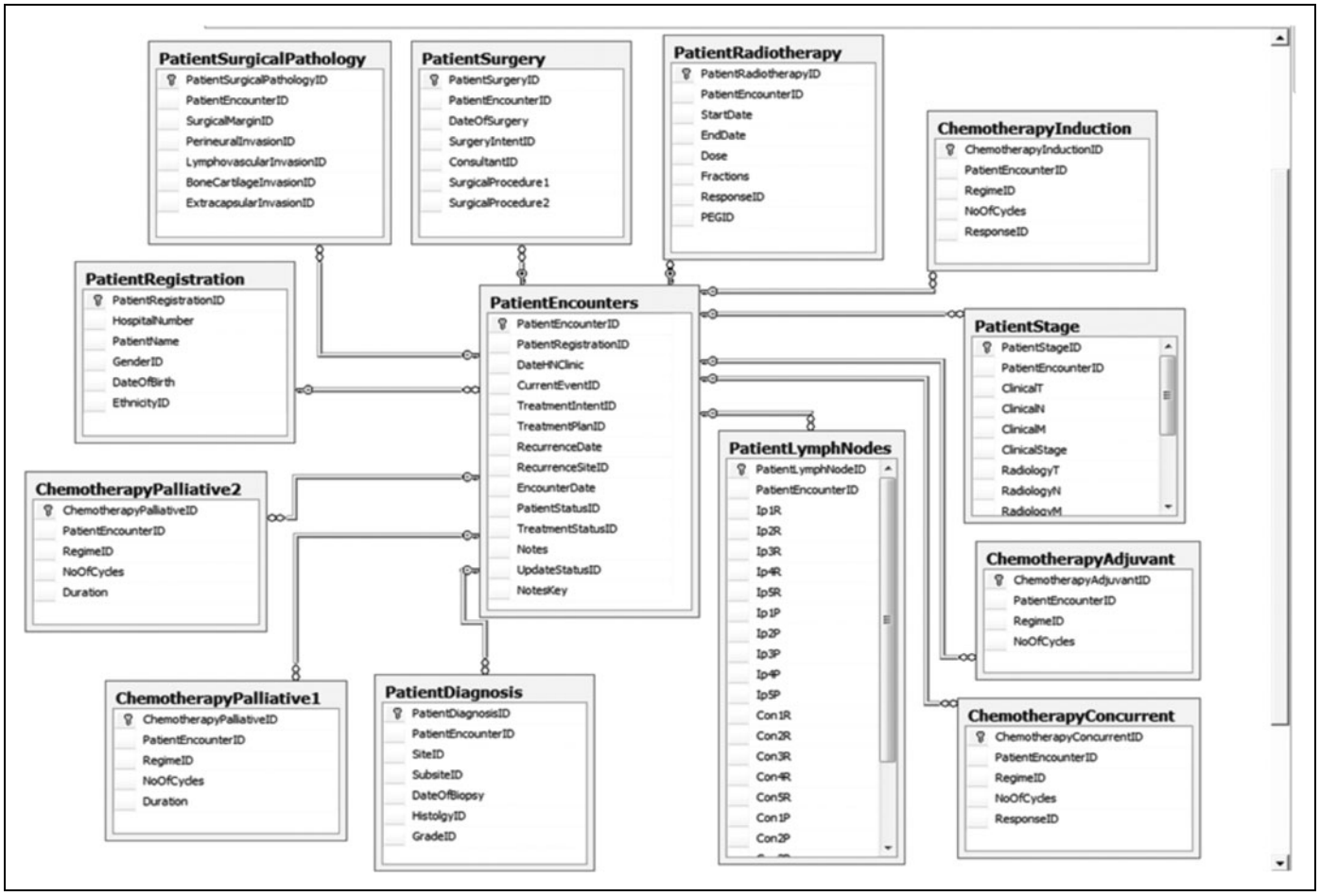

Fig. 5. IDB representation in the form of an entity relationship diagram. IDB, intermediate database.

designed to perform on the structured data in the IDB. For both descriptive and PA, the attribute "patient status" is used to represent the patient current status, such as "alive" and "dead."

Descriptive analysis. To perform a DA, we designed search queries for information retrieval from the IDB that have statistical significance for researchers. Using an SPSS tool, a researcher has the freedom to develop queries; however, every user is not capable of designing queries, especially those complex in nature. Based on the requirements of the researchers, we developed conditional queries in the form of filters in SPSS.

- How many men and women have survived HNC disease in the last x years, and how many of them are currently alive?
- How many HNC patients are included in each status (dead, on treatment, terminal, alive, etc.)?

The flexible design allows more challenging and complex queries to be added to the system based on the needs of the users.

Predictive analysis. The data are classified to predict the treatment plan for a new patient case. That is, it determines the probability that a patient will be given a certain treatment plan for the given conditions. We developed a survival analysis model on the basis of CART.[24,25] The specifications of the data and the development strategy of the CART model are described in *Figure 6*.

Similar to other machine learning models, development of the CART model involves prerequisite steps, including data preparation. Missing values and low-quality data degrade the performance and thus need to be removed from the data. Assuring the quality of data by removing incorrect values requires a lot of time; therefore, for this study, only oral cavity patients are considered for model development.

As described in *Figure 6*, the CART model has a number of merits over other classification methods. It is robust in the sense that it performs well even if its assumptions are somewhat violated by the model from which the data are generated. CART trees are simple to understand and interpret even by nonstatisticians.[24]

In the configuration of CART, we followed cross-validation, which is considered a stable validation method compared to split-sample validation. A default tree depth of five is

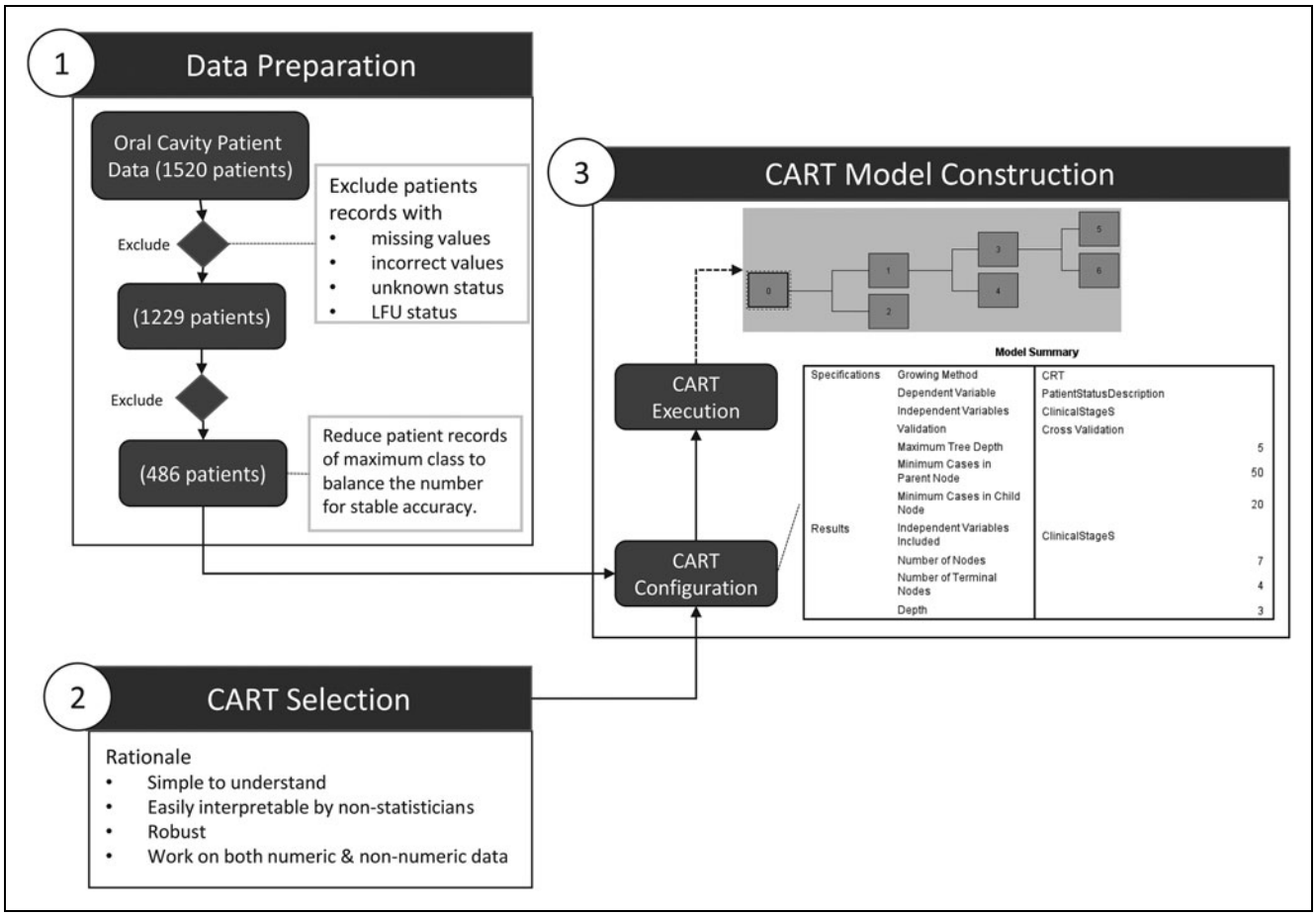

Fig. 6. CART model development process. CART, classification and regression tree. CRT, chemoradiotherapy; LFU, leave follow-up.

Table 3. Documents Parsing with Respect to Correct and Partially Correct Factors

DOCUMENT TYPE	TOTAL NO. OF DOCUMENTS	NO. OF CORRECTLY PARSED DOCUMENTS	NO. OF PARTIALLY CORRECT PARSED DOCUMENTS	CORRECTNESS RATIO (%)
Semistructured	14,341	14,197	144	99
Unstructured	469	454	15	97

considered for distributing cases involving 20 child nodes and 50 parent nodes. The model shown in *Figure 6* reflects the specifications for the survival analysis model on the basis of clinical staging. For survival models on the basis of other factors, the process remains the same except for the independent attributes. For instance, the survival analysis model on the basis of risk factors will involve the independent attributes related to risk factors such as smoking and alcohol.

Experiments and Results

EXPERIMENTAL SETUP AND ENVIRONMENT

The proposed SEAS is implemented in the vicinity of the HNC research center connected to the main hospital information system at SKMCH&RC. One oncologist and two resident doctors from SKMCH&RC participated in the analysis of data and the system. They delivered the knowledge of different information used in the clinical documents, which helps us as researchers to design and develop the system. They were also involved in validating the results. The SEAS receives the patient information from HMIS in the form of clinical documents. Overall, 18,621 clinical documents as deidentified copies are received by the SEAS for 3,811 patients. The information extracted by SEAS is checked and validated manually by the resident doctors and the results are compiled, which are presented in the following sections.

Table 4. Named Entity Recognition Results for 1,064 Lymphatic Attributes

LYMPHATIC ATTRIBUTES	TOTAL NO.	PRECISION	RECALL	F-MEASURE
Lymph node level i	264	0.96	0.97	0.97
Lymph node level ii	266	0.90	0.94	0.96
Lymph node level iii	248	0.93	0.95	0.96
Lymph node level iv	202	0.93	0.94	0.96
Lymph node level v	84	0.96	0.97	0.97
All attributes	1,064	0.97	0.94	0.96

EVALUATION CRITERIA

The evaluation criteria used for checking the correctness of converted semistructured and structured to structured format include completely correct parsing and partially correct parsing. The completely correct parsing criteria show the number of documents, of which every single data element is parsed correctly. While partially correct parsing represents the number of documents, of which one or more than one data elements are parsed incorrectly. NER performance is measured in the form of recall, precision, and F-measure using following formulas.

$$Precision(P) = \frac{TP}{TP+FP}, Recall(R) = \frac{TP}{TP+FN},$$
$$\text{and } F-measure\ (F) = \frac{2PR}{P+R}.$$

Results are provided in the top–down approach. First, the performance of IES is provided at the document parsing level (*Table 2*), and at a granular level results are provided only for unstructured documents. Second, the performance of SAS is provided for the methods of descriptive and PA.

INFORMATION EXTRACTION RESULTS

Document-level results: Out of 18,621 documents, 3,811 documents fall in the structured category, which are saved correctly to the IDB without any ambiguity. The maximum

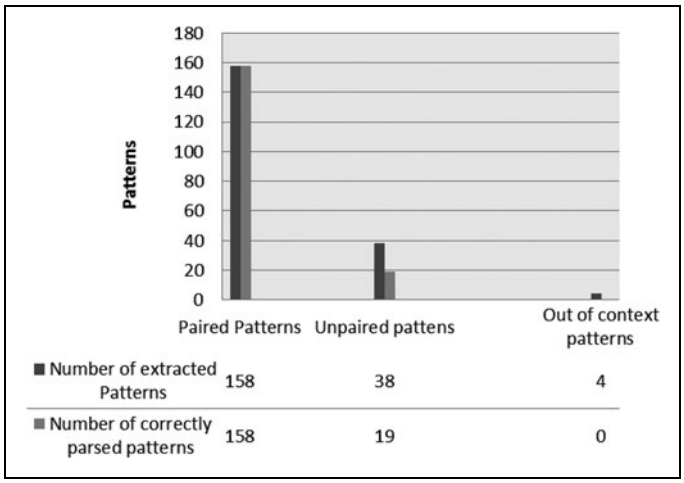

Fig. 7. Correctly parsed patterns of the total extracted patterns.

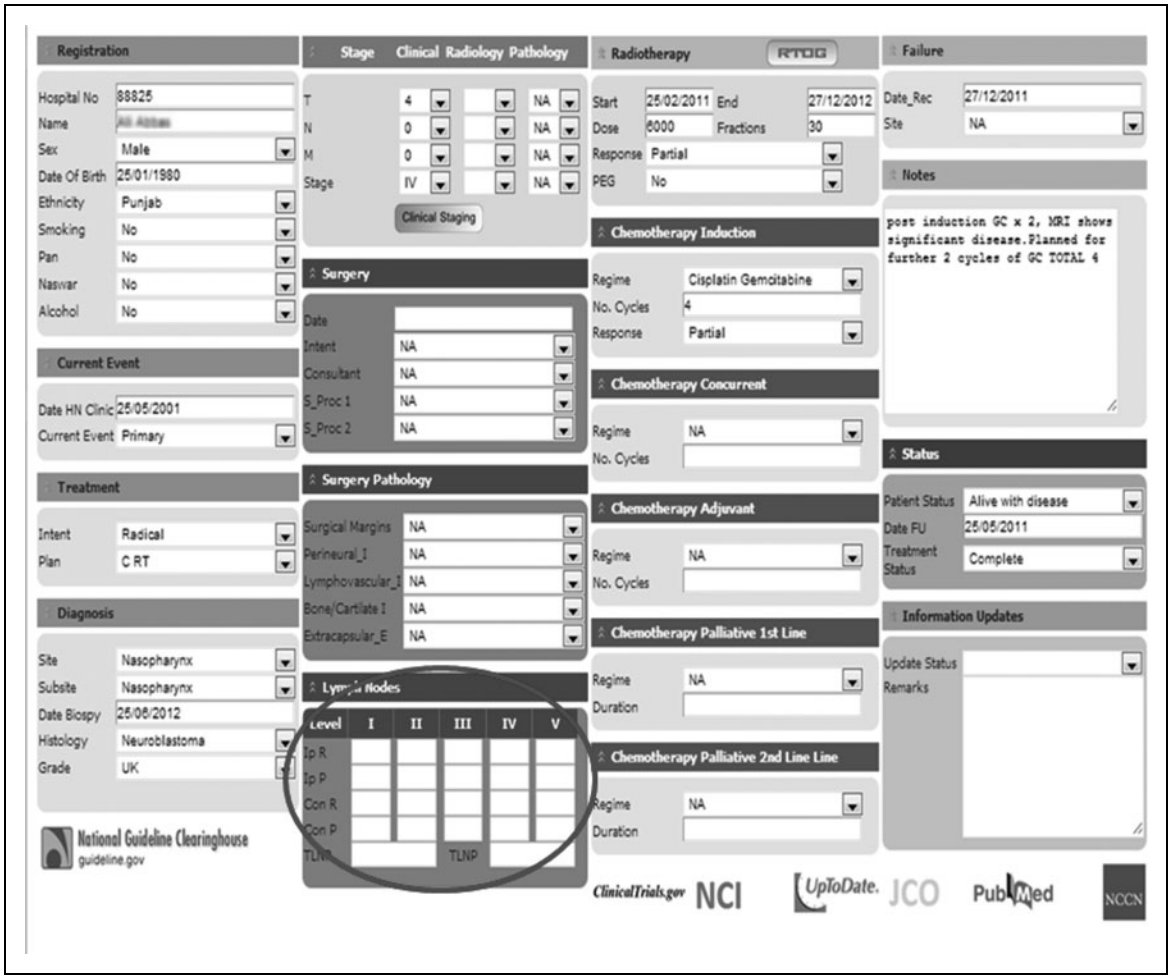

Fig. 8. GUI of implemented system with highlighted lymphatic section. GUI, graphical user interface.

number of documents falls in the category of semistructured, which are converted into the structured format with 99% accuracy. The unstructured documents, because of their complexity, are parsed with 97% accuracy.

Named entity recognition: In *Table 3*, we present results for histopathology notes, where we used permutation-based PR technique. Because of time-consuming activity, the evaluation is performed on a subset of (198 documents) unstructured histopathology notes that contained a total of 1,064 lymphatic attributes. Proposed NER correctly recognized named entities in histology reports by achieving an average F-score of 0.96. The proposed system overall results of NER with 1,064 lymphatic attributes are shown in *Table 4*.

PR results: Using permutation-based PR method, three types of patterns are found in the unstructured text.

- Paired patterns are the patterns with equal number of keys and values.
- Unpaired patterns show the patterns having unequal number of keys and values.
- Out-of-context patterns include the patterns out of the scope of 64 permutations described in Eq. (1).

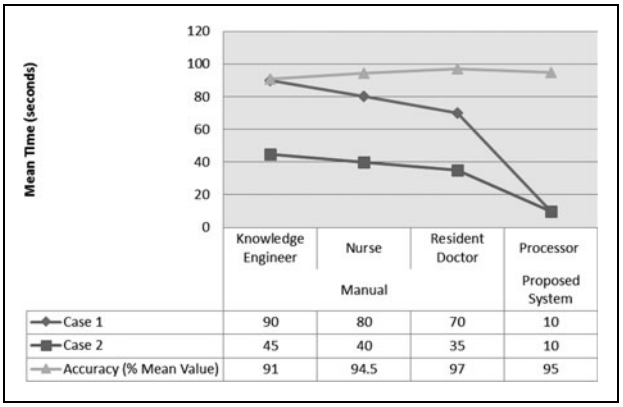

Fig. 9. Time (mean) comparison for manual and proposed systems.

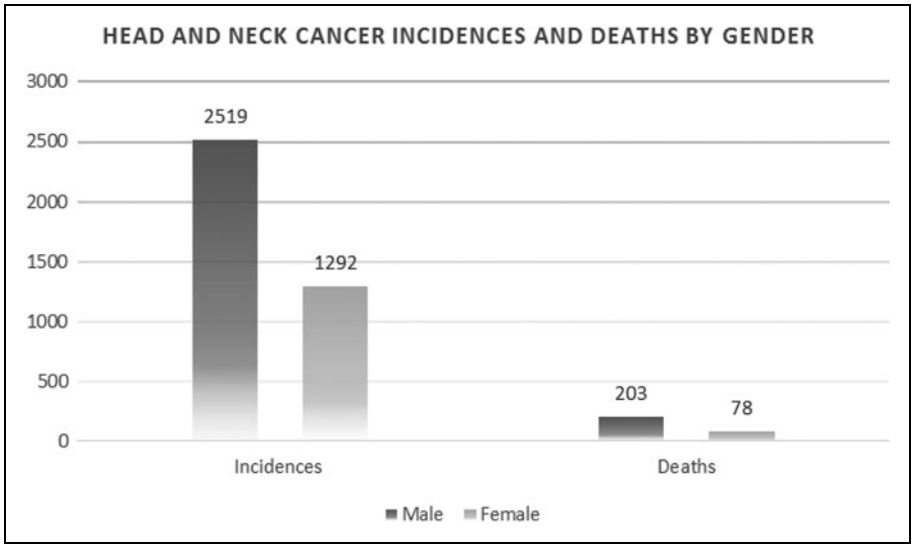

Fig. 10. Head and neck cancer incidences and death by gender.

As shown in *Figure 7*, the category of paired patterns produced better parsing results compared to the other two categories. Only 50% of the unpaired patterns are successfully resolved. The complete list of patterns discovered from narratives is given in Appendix.

Time and accuracy of the system. The overall objective of information extraction was to maximize clinical throughput by minimizing the manual efforts spent on data entry. Manual data entry is a laborious and time-consuming task. To measure the performance of the IES in terms of time, we examined the time spent on slot population of the most challenging part (lymphatic information) as highlighted in *Figure 8*. The time is recorded for 40 randomly selected documents created by three types of people with different levels of expertise: resident doctor, nurse, and knowledge engineer. In the manual system, all of the values for the slots are entered by a human manually, while the proposed system populates them automatically. The time and mean accuracy of the system are recorded for the following two cases:

- Case 1: *population of complete lymphatic information,*
- *Case 2: population of partial lymphatic information.*

The results for people with different levels of expertise were different with regard to both time and accuracy. Since resident doctors' expertise levels are high, they were able to populate the slots in less time and with higher accuracy. Knowledge engineers more slowly populated the slots with less accuracy compared to the other two groups due to their lower expertise level. In addition, we recorded the time and accuracy for automatic extraction by IES. Note that the time includes the time spent on input entry (patient number) and information extraction (lymphatic) and excludes the system setting time, which is required only once at the beginning. We calculated the time difference between manual slot population and automated slot population based on the two cases, as shown in *Figure 9*. The mean accuracy value was also calculated for the two cases in both the manual and proposed systems.

SAS RESULTS

The structured information in the IDB is a set of information grouped in different categories such as risk factors, diagnosis, staging, treatments (surgery, chemotherapy, and radiotherapy), status, and others. Among 3,811 patients, a total of 2,722 male and 1,370 female patients are found with HNC disease.

DA results: As shown in *Figure 10*, 203 deaths are recorded while other 2,519 are found survived in the male category. In the female category, 78 patients are observed dead, while other 1,292 are found in alive state.

Figure 11 provides the details of patient status, survived and dead. Alive state shows the patients who are cured successfully

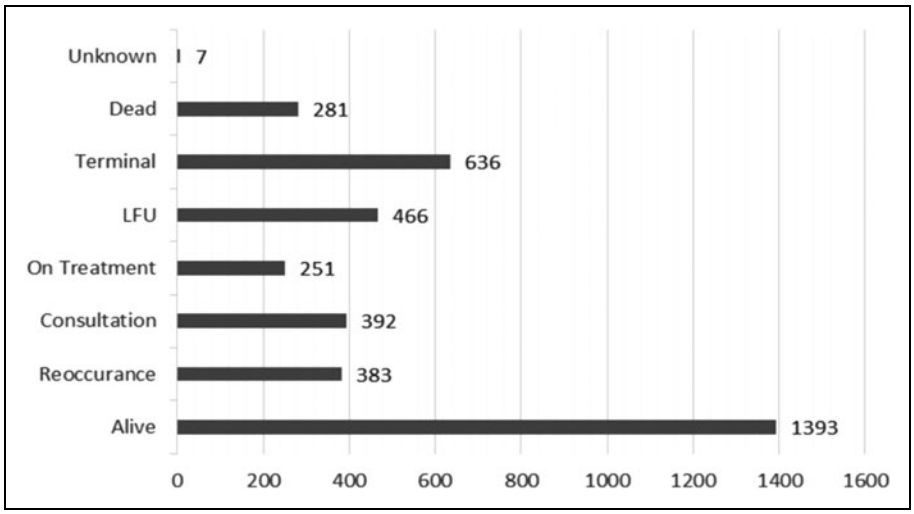

Fig. 11. Survivability ratio.

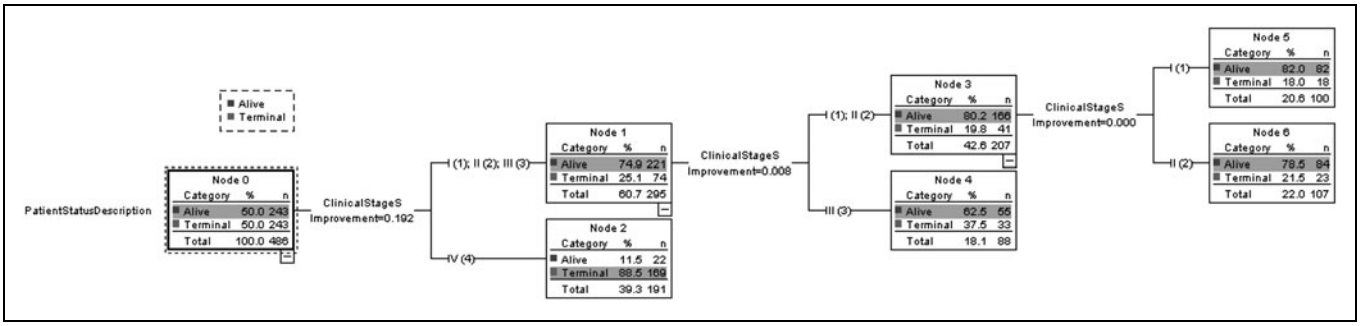

Fig. 12. Patient survival classification and prediction on the basis of clinical stage.

and they are in the highest number. Following the highest, terminal patients are at second, the patients who are alive with disease but are incurable. Leave follow-up are the patients who failed to maintain the follow-up with the hospital. Overall, 281 patients are observed dead among 3,811 patients.

Many other interesting descriptive statistical results are achieved from these data presented in different studies.[26–28]

PA results: The most interesting part of survival analysis is the PA, which is based on CART decision tree. For the development of CART model, a subset of data that consists of only oral cavity patients is utilized. Patient survival status is checked in correlation with clinical stage values and with lifestyle risk factors, smoking, Naswar (a moist, powdered tobacco snuff), pan, and alcohol values.

The CART model in *Figure 12* shows the results of patient survival status (alive, terminal) with respect to clinical stage. The most important survival analysis results are summarized as around 88% patients are in the terminal state if they are in the ClinicalStageS = 4. While patients in the ClinicalStageS = 1, 2, or 3 are observed mostly in the alive state (75%). The survival predictions derived from the results in *Figure 12* are described in *Table 5*.

Lifestyle risk factors, as shown in *Figure 13*, clearly indicate the patients with risk factor "Naswar = yes" are observed mostly (66%) in the terminal state. Among the 273 terminal state patients, 155 patients (about 64%) are the patients for whom there is at least one lifestyle risk factor as positive.

Table 5. Survival Chance of Patients on the Basis of Clinical Stage Values

CLINICAL STAGE VALUE	CHANCE OF SURVIVAL, %
I	82
II	78.5
III	62.5
IV	11.5

The survival predictions derived from the results in *Figure 13* are described in *Table 6*. It can be deduced from these results that chances of survival decrease for the patients who are addicted to lifestyle risk factors compared to nonaddicted cancer patients.

Discussion

The proposed system objectives include saving time (in terms of cost and effort) on the conversion of unstructured data to structured data. The system provides an automatic conversion and storage process for transforming data into a structured format. A benefit of the proposed system is the definition of services for the structured data such as survival analysis. Therefore, we assert that using the proposed system for the conversion of unstructured data to structured data saves time, reduces costs, and enables services to different stakeholders. We encountered challenges during the development of the proposed system that are discussed below.

Data analysis challenge: The information extraction process requires understanding the meaning of information from physicians for lexicon creation. These lexicons are used to recognize entities in the text. Therefore, an accurate analysis will result in accurate text recognition. As a prerequisite, we had face-to-face meetings, attended tutorials, and had informal sessions with physicians to better understand the meaning of information. Although we validated the extracted information from the physicians, there still exists some missing information due to the lack of domain understanding. A more suitable solution for the future is to provide an interactive graphical user interface for physicians to obtain information.

Data quality challenge: The predictive model requires quality data to perform better. Incorrect values in a structured format were another challenge faced during the design and

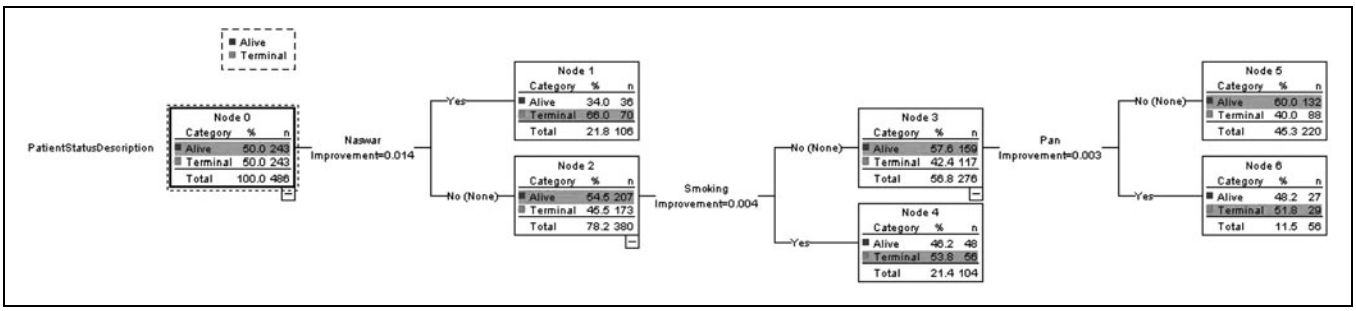

Fig. 13. Patient survival classification and prediction on the basis of risk factors.

development of the proposed system. For instance, the final clinical stage value is dependent on three values: tumor, node, and metastasis. In many records, the value is incorrectly recorded in the source document. During data preparation for the CART model, we manually checked all of the records and corrected the values. Furthermore, for some information fields, the values are recorded as "unknown"; these are removed from the data to keep only the known information to determine the prediction ability of the model. Such inconsistencies were handled at the time of data preparation to acquire quality data for prediction model development.

Limitations of the work: We highlight some of the limitations of the current work that we intend to overcome in future work. First, it is a challenging task to cover all types of unstructured documents. Although we introduced an innovative method of permutation-based PR for lymphatic information in histopathology reports, there is the possibility that the same patterns or patterns of a different nature will be found in chemotherapy or surgical notes. We plan to address this issue with a more comprehensive study to identify more patterns in unstructured documents. Another limitation of this work is related to the lexicon completeness. We developed the lexicon from the direct knowledge gained from domain experts and from the literature material associated with the existing system. A possible solution to this matter is to develop an independent subsystem for lexicon maintenance. In this way, the existing lexicon presented in the study can be grown to cover the maximum number of concepts in the HNC domain.

Table 6. Survival Chance of Patients on the Basis of Clinical Stage Values

RISK FACTORS	CHANCE OF SURVIVAL, %
Naswar = yes	34
Smoking = yes	46.5
Pan = yes	48.2

Future vision: The structured information created with the IES of SEAS can be grown into a big data repository in the future. Innovative big data technologies can be envisioned for application to big data analysis and research. Furthermore, it is feasible for other cancer research centers to replicate the IES and SAS methods to reduce the costs and the time spent on data formatting and analysis. The structured format repository will eventually be used with the big data technologies to generate analytics services that can be helpful in decision-making and recommendation systems.

Conclusion

Telemedicine and e-health envision smart services in smart environments. These smart services are dependent on data being in an understandable format. This study targeted a very common problem in today's e-health systems: the conversion of narrative data into a structured format. We proposed and developed the SEAS for clinical text extraction using an automated approach involving NLPreP, named entities, and PR. The SEAS reduced the time and energy of human resources spent unnecessarily on manual tasks. The techniques presented are extendable and ready to be replicated in other domains, for extracting and converting narrative text to a structured format. In the future, a self-learning mechanism feature can be added to the system to automatically update the domain lexicon and mentions. Furthermore, we plan to extend the system to other types of cancer, such as lung and breast cancers, after successfully passing the test phase for HNC.

Acknowledgments

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 Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2011-0030079).

Disclosure Statement

No competing financial interests exist.

REFERENCES

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics. *CA Cancer J Clin* **2010;**60:277–300.
2. Slavov V, Rao P, Paturi S, et al. A new tool for sharing and querying of clinical documents modeled using HL7 Version 3 standard. *Comput Methods Programs Biomed* **2013;**112:529–552.
3. Spasic I, Livsey J, Keane JA, Nenadic G. Text mining of cancer-related information: Review of current status and future directions. *Int J Med Inform* **2014;**83:605–623.
4. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin* **2011;**61:69–90.
5. Ricketts K, Williams M, Liu ZW, Gibson A. Automated estimation of disease recurrence in head and neck cancer using routine healthcare data. *Comput Methods Programs Biomed* **2014;**117:412–424.
6. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif Intell Med* **2005;** 34:113–127.
7. Liu F, Weng C, Yu H. Natural language processing, electronic health records, and clinical research. *Clinical Research Informatics*. Springer, **2012:**293–310.
8. Zhou X, Han H, Chankai I, Prestrud A, Brooks A. Approaches to text mining for clinical medical records. *Proceedings of the 2006 ACM symposium on Applied computing*. New York, NY, USA: ACM, **2006;**235–239.
9. Ananiadou S, McNaught J. *Text mining for biology and biomedicine*. Boston, London: Artech House, **2006**.
10. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb Med Inform* **2008;**35:128–144.
11. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study. *J Am Med Inform Assoc* **2008;**15:87–98.
12. Mykowiecka A, Marciniak M, Kupsc A. Rule-based information extraction from patients' clinical data. *J Biomed Inform* **2009;**42:923–936.
13. Vandenberghe HEE, Van Casteren V, Jonckheer P, et al. Collecting information on the quality of prescribing in primary care using semi-automatic data extraction from GPs' electronic medical records. *Int J Med Inform* **2005;**74:367–376.
14. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: A medication information extraction system for clinical narratives. *J Am Med Inform Assoc* **2010;**17:19–24.
15. Campbell DA, Johnson SB. A transformational-based learner for dependency grammars in discharge summaries. *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*. Stroudsburg, PA, USA: Association for Computational Linguistics, **2002:** 37–44.
16. Ruch P, Baud R, Geissbuhler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artif Intell Med* **2003;**29:169–184.
17. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. *Int J Med Inform* **2008;**77:81–97.
18. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: Towards better research applications and clinical care. *Nat Rev Genet* **2012;**13:395–405.
19. Sultan F, Aziz MT, Khokhar I, et al. Development of an in-house hospital information system in a hospital in Pakistan. *Int J Med Inform* **2014;**83:180–188.
20. Richardson L and Ruby S. *RESTful web services*. O'Reilly Media, Inc., Sebastopol, CA, 2008.
21. Denny JC, Miller RA, Johnson KB, Spickard A, 3rd. Development and evaluation of a clinical note section header terminology. *AMIA Annual Symposium proceedings*. Washington, DC: American Medical Informatics Association, **2008:**156–160.
22. Denny JC, Spickard A 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc* **2009;**16:806–815.
23. IBM. IBM SPSS Decision Trees 21. Available at www.sussex.ac.uk/its/pdfs/SPSS_Decision_Trees_21.pdf. 2012 (last accessed December 25, 2015).
24. Lewis RJ. An introduction to classification and regression tree (CART) analysis. *Annual meeting of the society for academic emergency medicine in San Francisco, California*. **2000:**1–14.
25. Rutkowski L, Jaworski M, Pietruczuk L, Duda P. The CART decision tree for mining data streams. *Inform Sci* **2014;**266:1–15.
26. Iqbal H, Bhatti AB, Raza Hussain AJ. Ten year experience with surgery and radiation in the management of malignant major salivary gland tumors. *Asian Pac J Cancer Prev* **2014;**15:2195–2199.
27. Iqbal H, Bhatti ABH, Hussain R, Jamshed A. Regional failures after selective neck dissection in previously untreated squamous cell carcinoma of oral cavity. *Int J Surg Oncol* **2014;**2014:205715.
28. Jamshed A, Hussain R, Iqbal H. Gemcitabine and cisplatin followed by chemo-radiation for advanced nasopharyngeal carcinoma. *Asian Pac J Cancer Prev* **2014;**15:899–904.

Address correspondence to
Sungyoung Lee
Ubiquitous Computing Lab
Department of Computer Science and Engineering
Kyung Hee University
Yongin 446-701
South Korea

E-mail: sylee@oslab.khu.ac.kr

Received: July 21, 2016
Revised: August 25, 2016
Accepted: August 28, 2016

(*Appendix follows*→)

Appendix

In Table A1, we present the patterns discovered from the unstructured text in the test document set of histopathology notes.

Table A1. Patterns Discovered from Real Lymph Node Narratives (Selected Set of 20 Histopathology Notes)

NO.	UNSTRUCTURED TEXT	DISCOVERED PATTERN
1.	LEVEL-1 LYMPH NODE, BIOPSY: Benign salivary gland tissue. Six lymph nodes, negative for tumor. LEVEL-5 LYMPH NODE, BIOPSY: 7 lymph nodes, negative for tumor. LEVEL-2A LYMPH NODE, BIOPSY: 9 lymph nodes, negative for tumor. LEVEL-2B LYMPH NODE, BIOPSY: 13 lymph nodes, negative for tumor. LEVEL-4 LYMPH NODE, BIOPSY: 25 lymph nodes, negative for tumor.	Level i 6 level v 7 level iiA 22 level iv 25
2.	LEFT SIDE OF NECK LEVEL-2B, NECK DISSECTION: 9 lymph nodes with no evidence of metastatic carcinoma.	Level iiB 8 level iiA 8 level iv 9 level i 4 level iii 3
	LEFT SIDE OF NECK LEVEL-2A, NECK DISSECTION: 8 lymph nodes with no evidence of metastatic carcinoma.	
	LEFT SIDE OF NECK LEVEL-4, NECK DISSECTION: 9 lymph nodes with no evidence of metastatic carcinoma. Unremarkable skeletal muscle.	
	LEFT SIDE OF NECK LEVEL-1, NECK DISSECTION: Unremarkable salivary gland, no evidence of malignancy.	
	Four lymph nodes with no evidence of metastatic carcinoma.	
	LEFT SIDE OF NECK LEVEL-3, NECK DISSECTION: 3 lymph nodes with no evidence of metastatic carcinoma.	
3.	LEFT NECK, LEVEL-1, NECK DISSECTION: Five reactive lymph nodes. Unremarkable submandibular gland.	Level i 5 level ii 6 level iii r 1 level iii p 2 level iv 8 level v 12
	LEFT NECK, LEVEL-2, NECK DISSECTION: Six reactive lymph nodes.	
	LEFT NECK, LEVEL-3, NECK DISSECTION: Metastatic carcinoma in 1 of 2 lymph nodes.	
	LEFT NECK, LEVEL-4, NECK DISSECTION: Eight reactive lymph nodes.	
	LEFT NECK, LEVEL-5, NECK DISSECTION: Twelve reactive lymph nodes.	
4.	RIGHT-SIDED NECK DISSECTION: 4 out of 30 lymph nodes positive for metastatic carcinoma (2/2 level-I, 1/5 level-II, 1/6 level-III, 0/8 level-IV, 0/9 level-V).	Ipsilateral: level i r 2 level i p 2 level ii r 1
		Level ii p 5 level iii r 1 level iii p 6 level iv r 0 level iv p 8 level v r 0 level v p 9
	LEFT-SIDED NECK DISSECTION: 1 out of 30 lymph nodes positive for metastatic carcinoma (1/8 level-I, 0/6 level-II, 0/6 level-III, 0/6 level-IV, 0/4 level-V).	Contralateral: level i r 1 level i p 8 level ii r 0 level ii p 6 level iii r 0 level iii p 6 level iv r 0 level iv p 6 level v r 0 level v p 4
5.	LEFT MANDIBLE LOWER BORDER, FROZEN SECTION:	0 level i r 3 level i p 0 level ii r 3 level ii p 1 level iii r 5 level iii p 0 level iv r 5 level iv p
	Positive for squamous cell carcinoma.	
	LEFT SIDE OF MANDIBLE, LEFT HEMIMANDIBULECTOMY WITH NECK DISSECTION: Squamous cell carcinoma, well differentiated, 2.0 cm. Tumor is infiltrating underlying bone. All soft tissue and bone resection margins are free of tumor. One out of 16 lymph nodes positive for metastatic carcinoma (0 out of 3 level I lymph nodes, 0 out of 3 level II lymph nodes, 1 out of 5 lymph nodes level III, 0 out of 5 lymph nodes level IV). Unremarkable salivary gland.	
6.	RIGHT SIDE OF NECK, LEVEL-1 NECK DISSECTION, BIOPSY: Metastatic tumor in 2 of 6 lymph nodes. Salivary gland, no evidence of tumor.	Level ii 8 level iii 18 level iv r 9 level iv p 10 level v r 7 level v p 12
	RIGHT SIDE OF NECK, LEVEL-2 NECK DISSECTION, BIOPSY: Metastatic tumor in all 8 lymph nodes.	
	RIGHT LEVEL-3 NECK DISSECTION, BIOPSY: Metastatic tumor in all 18 lymph nodes.	
	RIGHT LEVEL-4 NECK DISSECTION, BIOPSY: Metastatic tumor in 9 of 10 lymph nodes.	
	RIGHT LEVEL-5 NECK DISSECTION, BIOPSY: Metastatic tumor in 7 of 12 lymph nodes.	
	RIGHT PAROTID LYMPH NODE, BIOPSY: No evidence of tumor in 1 lymph node.	

continued →

Table A1. Patterns Discovered from Real Lymph Node Narratives (Selected Set of 20 Histopathology Notes) *continued*

NO.	UNSTRUCTURED TEXT	DISCOVERED PATTERN
7.	Right-sided lymph nodes reveal the following:-	Ipsilateral: level i r 2 level i p 6 level ii r 1 level ii p 6 level iii r 0 level iii p 2 level iv r 0 level iv p 3
	Level 1: 2 out of 6 lymph nodes show metastatic carcinoma.	
	Level 2: 1 out of 6 lymph nodes shows metastatic carcinoma.	
	Level 3: 0 out of 2 lymph nodes shows metastatic carcinoma. Level 4: 0 out of 3 lymph nodes shows metastatic carcinoma. Total right-sided lymph nodes: 3 out of 17 lymph nodes show metastatic carcinoma.	Contralateral: level i r 5 level i p 7 level ii r 0 level ii p 10 level iii r 0 level iii p 12 level iv r 0 level iv p 2
	Left-sided lymph nodes reveal the following:- Level 1: 5 out of 7 lymph nodes show metastatic carcinoma. Level 2: 0 out of 10 lymph nodes shows metastatic carcinoma. Level 3: 0 out of 12 lymph nodes shows metastatic carcinoma. Level 4: 0 out of 2 lymph nodes shows metastatic carcinoma. Total left-sided lymph nodes: 5 out of 31 lymph nodes show metastatic carcinoma.	
8.	LEFT LEVEL-2 LYMPH NODE, EXCISION BIOPSY: 1 lymph node shows reactive changes. RIGHT LEVEL-1 LYMPH NODE, EXCISION BIOPSY: 2 lymph nodes show reactive changes. LEFT LEVEL-3 LYMPH NODE, EXCISION BIOPSY: 9 lymph nodes show reactive changes. LEFT LEVEL-4 LYMPH NODE, EXCISION BIOPSY: 15 lymph nodes show reactive changes. LEFT LEVEL-5 LYMPH NODE, EXCISION BIOPSY: 2 out of 17 lymph nodes are positive for metastatic carcinoma. LEFT LEVEL-1 LYMPH NODE, EXCISION BIOPSY: 5 lymph nodes show reactive changes.	Level ii 1 level i 2 level iii 9 level iv 15
9.	LEVEL-III LYMPH NODE, RIGHT NECK DISSECTION: All 16 lymph nodes, negative for metastatic carcinoma.	Level iii 16 level iiA 0 level iiB 14
	LEVEL-IIA LYMPH NODE, RIGHT NECK DISSECTION: Fibroadipose tissue only. No lymph nodes identified.	
	LEVEL-IIB LYMPH NODE, RIGHT NECK DISSECTION: All 14 lymph nodes, negative for metastatic carcinoma.	
	LEVEL-I LYMPH NODE, RIGHT NECK DISSECTION: Salivary gland, unremarkable.	
	TONGUE, RIGHT PARTIAL GLOSSECTOMY: Moderately differentiated squamous cell carcinoma, 1.5 cm. Depth of invasion is 1.2 cm. All resection margins, free of tumor. No lymphovascular and perineural invasion seen.	
	LEVEL-IV LYMPH NODE, RIGHT NECK DISSECTION: 10 lymph nodes, negative for metastatic carcinoma.	
10.	LEVEL-I LEFT NECK, DISSECTION: 4 reactive lymph nodes, negative for metastatic carcinoma. Unremarkable submandibular salivary gland. LEVEL-II LEFT NECK, DISSECTION: 2 out of 13 lymph nodes positive for metastatic carcinoma with extracapsular effort. LEVEL-III LEFT NECK, DISSECTION: 4 out of 12 lymph nodes positive for metastatic carcinoma with extracapsular effort.	Level i 4 level ii r 2 level ii p 13 level iii r 4 level iii p 12
	TONGUE, LEFT PARTIAL GLOSSECTOMY: Moderately differentiated squamous cell carcinoma, 2.5 cm. Maximum depth of invasion, 1.0 cm. All resection margins, free of tumor (0.2 cm from closest anterior resection margin).	
11.	LEVEL-IIA, RIGHT NECK DISSECTION: 1 out of 19 lymph nodes, positive for metastatic carcinoma. LEVEL-I, RIGHT NECK DISSECTION: Salivary gland, free of tumor. 10 lymph nodes, free of tumor. LEVEL-IIB, RIGHT NECK DISSECTION: 13 lymph nodes, free of tumor. LEVEL-III, RIGHT NECK DISSECTION: 7 lymph nodes, free of tumor.	Level i 10 level ii 13 level iii 7
	RIGHT LATERAL TONGUE, HEMIGLOSSECTOMY: Moderately differentiated squamous cell carcinoma, 1.5 cm. Maximum depth of invasion is 0.4 cm. All resection margins, free of tumor. No perineural invasion seen.	
12.	LEVEL-IV, LYMPH NODES, BIOPSY: 9 lymph nodes, negative for metastatic carcinoma. LEVEL-IIA, LYMPH NODES, BIOPSY: 8 lymph nodes, negative for metastatic carcinoma. LEVEL-IIB, LYMPH NODES, BIOPSY: 10 lymph nodes, negative for metastatic carcinoma. LEVEL-III, LYMPH NODES, BIOPSY: 11 lymph nodes, negative for metastatic carcinoma.	Level iv 9 level iiA 8 level iiB 10 level iii 11
13.	LEFT CERVICAL LYMPH NODES, LEVEL 1–4, LEFT NECK DISSECTION: Residual squamous cell carcinoma in soft tissue neck. Salivary gland free of tumor. All 25 lymph nodes negative for tumor. Six out of 6 level-I lymph nodes free of tumor. Six out of 6 level-II lymph nodes free of tumor. Nine out of 9 level-III lymph nodes free of tumor. Four out of 4 level-IV lymph nodes free of tumor.	Level i r 6 level i p 6 level ii r 6 level ii p 6 level iii r 9 level iii p 9
14.	MANDIBLE, LEFT HEMIMANDIBULECTOMY: Squamous cell carcinoma poorly differentiated, 3.0 cm extending up to the level of minor salivary gland but not invading the parenchyma. All margins including bone are free of tumor.	Ipsilateral: level i 10 level ii 7
	SOFT TISSUE FROM NECK BILATERAL, RADICAL NECK DISSECTION:	Contralateral: level i 6 level ii 11 level iii 9 level iv 9
	RIGHT SIDE OF NECK: Level-I lymph node: 10 lymph nodes free of tumor. Level-II lymph node: 7 lymph nodes free of tumor. Salivary gland free of tumor.	
	LEFT SIDE OF NECK: Level-I: 6 lymph nodes free of tumor. Level-II: 11 lymph nodes free of tumor. Level-III: 9 lymph nodes free of tumor. Level-IV: 9 lymph nodes free of tumor. Salivary gland free of tumor. All 52 lymph nodes are free of tumor.	

continued →

Table A1. Patterns Discovered from Real Lymph Node Narratives (Selected Set of 20 Histopathology Notes) *continued*

NO.	UNSTRUCTURED TEXT	DISCOVERED PATTERN
15.	LEVEL-1 LYMPH NODES, BIOPSY: No evidence of malignancy in 5 lymph nodes. LEVEL-2 LYMPH NODES, BIOPSY: Metastatic squamous cell carcinoma in 1 of 7 lymph nodes. LEVEL-3 LYMPH NODES, BIOPSY: Metastatic squamous cell carcinoma in 2 of 8 lymph nodes. LEVEL-4 LYMPH NODES, BIOPSY: No evidence of malignancy in 15 lymph nodes. LEVEL-5 LYMPH NODES, BIOPSY: No evidence of malignancy in 10 lymph nodes. LEVEL-1 LUMP, LEFT SIDE OF NECK, BIOPSY: No evidence of malignancy in 3 lymph nodes (see note).	Ipsilateral: level I 5 level ii r 1 level ii p 7 level iii r 2 level iii p 8 level iv 15 level v 10
		Contralateral: level i
16.	LEFT SIDE OF TONGUE, PARTIAL GLOSSECTOMY: Well-differentiated squamous cell carcinoma. All resection margins free of tumor. LEVEL-1 LYMPH NODE, BIOPSY: Benign salivary gland tissue. Six lymph nodes, negative for tumor.	Level i 6 level v 7 level iiA 9 level iiB 13 level 4
	LEVEL-5 LYMPH NODE, BIOPSY: 7 lymph nodes, negative for tumor. LEVEL-2A LYMPH NODE, BIOPSY: 9 lymph nodes, negative for tumor. LEVEL-2B LYMPH NODE, BIOPSY: 13 lymph nodes, negative for tumor. LEVEL-4 LYMPH NODE, BIOPSY: 25 lymph nodes, negative for tumor.	
17.	RIGHT SIDE NECK DISSECTION: 3 out of 26 lymph nodes show metastatic carcinoma: Level 1: 3 out of 4 lymph nodes positive for metastatic carcinoma. Level 2: All 2 lymph nodes, free of tumor. Level 3: All 12 lymph nodes, free of tumor.	Ipsilateral: level I r 3 level I p 4 level ii 2 level iii 12 level iv 8
	Level 4: All 8 lymph nodes, free of tumor. Submandibular gland, free of tumor.	Contralateral: level ii r 1 level ii p 9 level iii 9
	LEFT SIDE OF NECK, LYMPH NODE, DISSECTION: 1 out of 18 lymph nodes shows metastatic carcinoma. Level 2: 1 out of 9 lymph nodes positive for metastatic carcinoma. Level 3: All 9 lymph nodes and submandibular gland free of tumor.	
	TONGUE, HEMIGLOSSECTOMY: Poorly differentiated carcinoma. All resection margins free of tumor.	
18.	LYMPHATICS, LEFT SIDE NECK, RADICAL NECK DISSECTION: 10 out of 23 lymph nodes are positive for metastatic carcinoma (4 out of 4 level one lymph nodes, 0 out of 1 level two lymph nodes, 3 out of 7 level three lymph nodes, 0 out of 6 level four lymph nodes, and 3 out of 5 level five lymph nodes).	4 level I r 4 level I p 0 level ii r 1 level ii p 3 level iii r 7 level iii p 0 level iv r 6 level iv p 3 level v r 5 level v p
19.	LEFT LEVEL-2 LYMPH NODES, BIOPSY: 1 of 6 lymph nodes with metastatic carcinoma. LEFT NECK, LEVEL-1 LYMPH NODES, BIOPSY: 4 lymph nodes, no tumor found.	Level ii r 1 level ii p 6 level i 4 level iv 6 level v 14
	LEFT NECK, LEVEL-4 LYMPH NODES, BIOPSY: 6 lymph nodes, no tumor found. LEFT NECK, LEVEL-5 LYMPH NODES, BIOPSY: 14 lymph nodes, no tumor found.	
	LEVEL-3 LYMPH NODES, BIOPSY: 9 lymph nodes, no tumor found.	
20.	MID LINGUAL MARGIN, FROZEN SECTION Squamous lined mucosa containing inflamed granulation tissue. There is no evidence of malignancy.	Level i r 1 level i p 8 level ii r 0 level ii p 7 level iii r 0 level iii p 7 level iv
	MID BUCCAL MARGIN, FROZEN SECTION: Unremarkable squamous mucosa. There is no evidence of malignancy.	
	RIGHT MANDIBLE MANDIBULECTOMY: Well-differentiated squamous cell carcinoma, 2.0 cm. All resection margins are negative (closest superior margin 0.5 cm away). Metastatic carcinoma in 6 of 29 lymph nodes (5/7 Level I, 1/8 Level II, 0/7 Level III, 0/7 Level IV).	