International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

SNOMED CT Annotation for Improved Pathological Decisions in Breast Cancer Domain

G. Johanna Johnsi Rani, Gladis D, Joy John Mammen

Abstract: ***Breast cancer pathology reports are used in the diagnosis of the disease and determination of the stage of cancer in a patient. These reports are written or electronically generated by the Pathologist in English. The contents of a Pathology report generated by the Pathologist are usually in unstructured natural language form. The contents of a report are used to determine the Pathological classification and Cancer stage of a patient. Information extraction and making pathological decisions from natural language text is a complex process due to the heterogeneity of the report structure and its contents. The reports can be homogenized using the global annotation standard Systematized Nomenclature of Medicine – Clinical Terms, SNOMED-CT. It enables consistent representations of medical terms and can be used for clinical decision support systems (CDSS) and cancer reporting. SNOMED is a vast repository and its enormity and complexity necessitates extraction of a subset for a particular domain before using it for annotation. The annotation is performed either in online mode at the time of generation of the report or in offline mode on a batch of archived reports. A CDSS prototype is developed for breast cancer domain, which provides support to the Pathologist to determine the Pathological Classification and Cancer Staging on both natural language text and SNOMED-annotated text. With regard to Pathological decisions, a hypothesis is formulated that Annotation using SNOMED does not improve the system's performance in determining the cancer stage of a patient. For annotating the text, the system initially extracts a SNOMED subset for the domain. Performance Analysis of the decision support processes was done by determining Precision, Recall, Specificity, Accuracy, F-measure and Error. The analysis indicates that the annotation feature improved the accuracy of automated Pathological decisions presented by the CDSS to the Clinician for finalizing his decisions. In the future, the CDSS feature can be applied to other cancer domains and thus provide a means to improve decision-making related to those domains.***

Keywords**: Clinical Decision support system, SNOMED-CT, Natural Language processing, Information Extraction, Breast cancer pathology**

I. INTRODUCTION

India is ranked at the top in breast cancer deaths among women. Clinical decision support systems that provide decision support to the Medical practitioners can assist and speed-up the diagnosis, treatment and follow-ups in the treatment of diseases. To provide a decision-support tool to the Pathologists, in breast cancer domain, a prototype is designed and developed with essential support services. One of the services is provided for performing Pathological classification and cancer staging for an individual patient, using the Pathology report. Pathology reports are electronic documents in English language that provide pathological details of the patient. A Pathologist performs Pathological classification and cancer staging manually by examining the contents of the report and using the American Joint Committee on Cancer (AJCC) guidelines. Clinical Decision Support Systems that provide assistance to the Pathologist to perform Pathological analysis on patients must do so with high precision by processing unstructured data in the Pathology report. Information extraction using pattern-matching and Natural Language Processing steps help the data to be brought to a structured form. Further, a Pathology report written or electronically generated by the Pathologist is highly heterogeneous in structure and language. Converting the medical terms in the report to a standard form and annotating them would support the information extraction and subsequent pathological processes to be performed by the system. There are several medical codes for representation and storage of health information, a few important ones among them include the International Classification of Diseases (ICD) of the World Health Organization [1], Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) published by SNOMED International [2] and Logical Observation Identifiers, Names and Codes (LOINC), which is primarily used in medical laboratories [3]. The standardization process on Pathology reports can be performed using indigenous guidelines for individual hospitals or using global standards. Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) maintained by an International community is a global annotation standard that provides standardized terminology for numerous health care domains, with multilingual support. The Electronic Health Record (EHR) Standards for India released by the Ministry of Health & Family Welfare, Government of India, recommends SNOMED CT as one of the recommended Healthcare IT standards [4].

Manuscript published on 30 September 2019
* Correspondence Author
G. Johanna Johnsi Rani*, Department of Computer Science, Madras Christian College (Autonomous), University of Madras, Chennai, India. Email: johanna.g@mcc.edu.in
Gladis D, Principal, Bharathi Women's College (Autonomous), University of Madras, Chennai, India, Email: gladischristopher@gmail.com
Joy John Mammen, Professor and Head, Department of Transfusion Medicine & Immunohaematology, Christian Medical College, Vellore, India, Email: : joymammen@cmcvellore.ac.in

SNOMED is a huge repository of medical terminologies with more than 3,90,000 concepts, 11,60,000 English descriptions and 1.40 million relationships. The Pathological decision support requires preprocessing of report contents, SNOMED annotation and Information extraction for the Pathological classification and cancer staging. Preprocessing is done applying NLP tasks but due to the enormity of SNOMED database, extraction of a subset for the disease domain is essential as a preprocess to annotation. SNOMED was made available in India from 2014 and the developed CDSS extracts and uses breast cancer-specific SNOMED subset for annotation of reports that is used to improve the Pathological analysis process.

The need and benefits of using SNOMED CT in EHRs is manifold. It can serve as a common language that makes the information understandable and usable internationally. Using SNOMED annotation, the heterogeneous natural language text can be standardized and transmitted without loss of meaning to both human and machine understanding. The homogenization process through annotation also results in improved retrieval and interpretation by automated health systems. It enhances communication and electronic exchange of medical information. SNOMED also allows natural language processing to perform complex queries, and can cross-map to other medical terminologies such as International Coding of Diseases (ICD).

Considering the fact that Breast cancer is the number one killer disease among women in India, a CDSS prototype for breast cancer pathology domain is developed that automates two vital pathological decision-making processes namely Pathological classification and cancer staging. The Pathological Classification pTNM determines the classification of Tumour pT, Lymph node pN and Distant Metastasis pM. Grouping of these parameters determine the cancer stage of a patient. The CDSS permits the Pathologist to determine the pTNM and cancer stage both on natural language text and SNOMED-annotated text. Comparative analysis of the two methods of Pathological diagnosis is done by the CDSS. The dataset used for the work are breast cancer pathology reports obtained from a hospital in South India.

The remaining sections of this paper are organized as follows: Section II describes Related Works on Clinical Decision Support Systems in breast cancer domain with SNOMED CD support, Natural Language processing, SNOMED subset extraction and annotation. Section III explains the Materials and Methods used in SNOMED subset extraction, Annotation of reports and Cancer staging on Pathology reports in English and SNOMED-annotated reports, Section IV presents the Results and comparative analysis of the cancer staging on natural language text and annotated text and Section V presents the Conclusion.

II. RELATED WORKS

Many countries of the world use Electronic Health Records and various research works and CDSS developments have been carried out. A Web clinical decision support system for clinical oncologists that uses the characteristics of the individual patient and for patients making prognostic assessments is developed by Ana Fernandes et. al. [5] DESIREE is a European-funded project has developed three decision support systems for breast cancer domain: a guideline-based, an experience-based, and a case-based DSSs, that operate simultaneously and offer multimodal decision support to clinicians. [6] Research scholar Wanyonyi Peter Simeon, developed a prototype Decision Support System for the Diagnosis of Breast Cancer using Fuzzy Logic and Case Based Reasoning [7] R. K. Sinha et. al proposed a Novel Knowledge Based Decision Support System model in breast cancer treatment with five modules namely Patient Information Module, Search Engine Module, Knowledge base module, Case base module and Statistical module for improving the accessibility and efficient use of domain specific clinical data. [8]

In the Indian Subcontinent, the use of EHRs and CDSS is minimal until now and is used in a very small percentage of hospitals. A study was performed to know the perception of CDSS in Indian health care industry and it revealed that India lags behind in health IT when compared to other countries. Even the awareness level about CDSS was found to be low, with administrative staff being least aware of CDSS, followed by doctors and none of the medical students were aware of CDSS. [9] In a survey conducted by the authors among doctors in India, 75% of the respondents did not know about CDSS and indicated that their hospitals do not use CDSS. All the respondents of the survey indicated that they require CDSS for their profession. Studies indicate that India is yet to put its first step forward in advocating the use of EHRs in all the hospitals in the country and adapting CDSSs In order to improve the quality of patient care, awareness about CDSS must be created in India and the Indian healthcare system must develop CDSSs and popularize the use of CDSS in hospitals across the country.

CDSS that uses SNOMED-CT, a work similar to the developed CDSS was done on preventive care by Bader Al-Hablani. In assessing the use of SNOMED CT in CDSS and analyzing its if it improved preventive care, it was inferred that SNOMED reduced medical errors and improved preventive care. [10] Information extraction and Cancer staging on Natural Language text Processing were done earlier. [11, 12] To process natural language text, Production rules, Semantic categories, First-order Logic and Bayesian and Semantic networks [13], Symbolic rule-based classification methodology, [14, 15] and Pattern-based extraction [16] were performed. The cancer staging in the developed CDSS is done using American Joint Committee on Cancer (AJCC) guidelines. [17]. Pathology reports were used in the development of web-based search applications [18], and also converted to machine readable form using NLP [19]

In many countries outside India, Clinical reports are coded using medical coding systems (SNOMED / ICD) and UMLS Knowledge resources. SNOMED had been in real-time usage for many years in countries abroad. Only in March 2014, India became a Member, joining the global effort to develop, maintain, and enable the use of SNOMED CT terminology in health systems. SNOMED subset extraction for breast cancer domain was done using a fixed set of queries to identify concepts relating to a disease-specific subdomain Krystyna Milian et al. used formal clinical guidelines to define a subdomain as a set of concepts without considering relevance between concepts.

Zharko Aleksovski and Merlijn Sevenster included the aspect of relevance by applying the "Term frequency-Inverse document frequency" (TFIDF) measuring scheme. [20, 21] Rodríguez-Solano C, Cáceres J, and Sicilia M.A., developed a system that automatically generated subsets by traversing SNOMED relationships using glossary terms in clinical guidelines as seed terms. [22]. In earlier works on CDSS, CDSS for breast cancer domain provided support for breast cancer screening [23], diagnosis [24, 25] using a conjunction of clinical and pathological data with Genomic tools for treatment decisions [26] raising user alerts, symptom checking and risk stratification and providing treatment recommendations. SNOMED annotations were performed on natural language text with NLP and they were found to improve the semantic interoperability of documents. [27, 28, 29]

III. MATERIALS AND METHODS

Indigenous data was used in this work to determine cancer stage of patients. The text corpus used in this work is from a hospital in the region consisting of 150 de-identified breast cancer pathology reports. in .pdf or .txt format.

1) 18462/12
SPECIMEN : Right MRM
CLINICAL : Carcinoma right breast.
GROSS : Specimen of right modified radical mastectomy measuring 19x11x7cm with nipple bearing skin measuring 14x9cm and attached axillary pad of fat measuring 5.5x4x1cm. The nipple and areola appears to be uninvolved. Sectioning reveals a tumour measuring 4x3.5x3.5cm in the outer upper quadrant extending into the central region and is 1cm away from the nearest deep resection margin. It is hard with greyish white cut surface and is surrounded by dense fibrosis. The rest of the breast appears predominantly fibrofatty. Sectioning the axillary pad of fat reveals 5 lymph nodes, largest measuring 1.3x1x0.8cm with a greyish brown cut surface.
A) Nipple and areola 5 all (A1-A5)
B) Tumour 5 bits (B1-B5; B1 tumour with nearest deep resection margin)
C) Upper outer quadrant 1 bit (1 block)
D) Lower outer quadrant 1 bit (1 block)
E) Lower inner quadrant 1 bit (1 block)
F) Upper inner quadrant 1 bit (1 block)(B-F; The deep resection margin inked with india ink.)
G) 5 axillary lymph nodes 10 all (G1-G6; G1 and G2 from the same node)
NS/lv
MICRO :
A) Shows skin of nipple and areola, free of tumour.
B) Shows breast parenchyma infiltrated by a tumour arranged in tubules, nests, clusters and trabeculae displaying moderate pleomorphism, coarse nuclear chromatin, inconspicuous nucleoli and amphophilic cytoplasm. High grade DCIS is present. Lymphovascular invasion is evident. Perineurial invasion is not evident. Tumour is 1 cm from deep resection margin.
C-F) Shows breast parenchyma with no specific lesion.
G) Shows 5 lymph nodes, free of tumour.
IMPRESSION : Invasive ductal carcinoma, grade-II, right modified radical mastectomy.
Maximum size of the tumour is 4cm.
High grade DCIS present.
Lymphovascular invasion present.
Perineurial invasion, not evident.
Tumour is 1cm from the deep resection margin.
Skin of nipple and areola, free of tumour.
5 lymph nodes, free of tumour.
pT2N0Mx.

Fig. 1 Pathology report of a Patient

One of the main processes of the CDSS are Pathological classification and Cancer Staging. The developed CDSS prototype performs several preprocessing steps on the report contents before the main processes namely NLP based preprocessing of contents in the reports and SNOMED subset extraction. In the first preprocessing, the contents of the reports are standardized for easy retrieval and annotation. The 11 preprocessing steps are listed below: *Report segregation* in which multiple reports are separated into individual reports; *Section segmentation* in which the contents of a report are divided into its constituent sections; *Standardization of measures* in which all measures are converted to a uniform measure. For example, tumour sizes are either given in centimeters or millimeters and this step converts all the measures into millimeters; *Date preprocessing* in which all dates are converted to a uniform DD/MM/YYYY format; *Sentence segmentation* in which the contents of each section are separated into individual sentences. Period (.) is used to identify the sentences, with handling of exceptions for fraction values; *Standardization of numerical values*: The pathology reports have numeric values represented in numerals (3), or in English words (three). Such numerical values are standardized to Arabic numerals; *Alpha numeric representations*: The number of lymph nodes are represented as '1/3', or '1 out of three', or 'one out of three'. This value is converted into complete textual form as 'one out of three'; *Abbreviations*: Abbreviations are expanded by the system; *Spelling variations*: All discrepancies in spelling between British and American English are standardized using British English; *Whitespace removal*: The whitespaces are removed from the document. This improves the data extraction process; *Handling parenthesized terms*: Parentheses () or [ ] in the document are homogenized into [ ]. *Case sensitivity*: All text comparisons are made by converting the terms into lower case. In case of medical terms such as Ductal Carcinoma in situ, the terms are converted to a form as found in SNOMED. *Missing headers*: The pre-processing module appends missing headers into the document whenever necessary.

SNOMED is a comprehensive repository of medical terms relating to numerous diseases, their symptoms, the diagnostics factors, the procedures for diagnosis, classification of various stages of a disease and many others. The second important preprocessing to be performed before the annotation of reports using SNOMED is the extraction of SNOMED subset for the breast cancer pathology domain. A survey on SNOMED implementations identified two types of SNOMED subsets namely Data entry subsets and Data retrieval subsets. Data entry subsets have concepts of a specific disease domain and are used for recording patient encounters. Data retrieval subsets are a collection of concepts retrieved from SNOMED encoded data. [27] The subset extraction performed in the work presented is of the Data entry type. The need for SNOMED subset extraction is justified because it is sufficient for the annotation process to have only information relating to breast cancer, the disease under consideration. The extracted SNOMED subset also would reduce the size of database to be used in an application, thus increasing the speed of processes. The workflow of the CDSS rule engine that extracts SNOMED subset for annotation is shown in Figure 2.

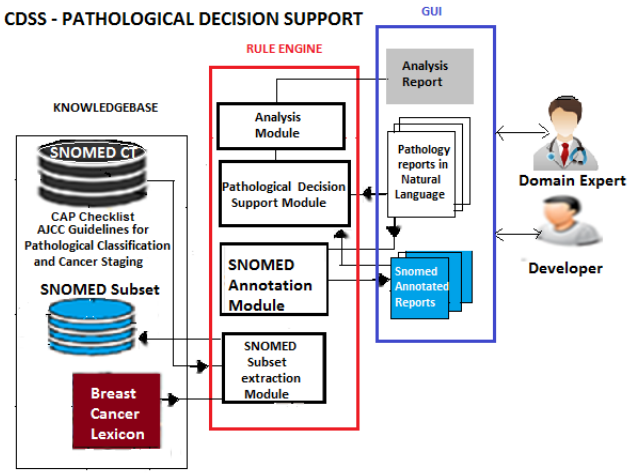


Fig. 2 CDSS Workflow for Pathological Decision Support

The first method of SNOMED subset extraction is Query-based, in which list of key-terms appended with the term "Breast", are used in the queries to extract the tuples from the SNOMED database.

The keyterms are: cancer, carcinoma, tumour, tumor, lymph node, metastatic, pT, pN, and pM, pTNM, stage, pathological, specimen, procedure, malignant, benign, mastectomy and neoplasm. The algorithm that is used for Subset extraction is given below:

i. Use SNOMED database
ii. SNOMED Subset S = φ
iii. Extract the tuples set s, relating to 'breast'
iv. Delete tuples relating to 'Male'
v. Append the other tuples S = S ∪ s
vi. For Key_term = 1 to n
Apply filtering on the Description table
Eliminate duplicates from S
Update to S

In the SNOMED subset, we eliminate the tuples relating to "Male breast". With every query-based extraction with a new key-term, the number of terms extracted and the number of new terms added for a search are updated. This helps in identifying the key-terms that generate maximum number of tuples for the SNOMED CT subset. The system generates a summary of the number of tuples extracted when each key-term is applied, and the total number of terms extracted. The second method of SNOMED subset extraction is performed using a Lexicon of breast cancer pathology terms. The Lexicon consists of key terms relating to the Breast cancer pathology domain which are automatically compiled from the dataset and the breast cancer standards namely AJCC protocol and College of American Pathologists (CAP) checklist. These Lexicon terms are used as search key terms to extract the SNOMED subset. Fig. 3 shows the tuples extracted from the SNOMED database using a query with search term 'breast'. A total of 4859 tuples were extracted with key term 'breast'.

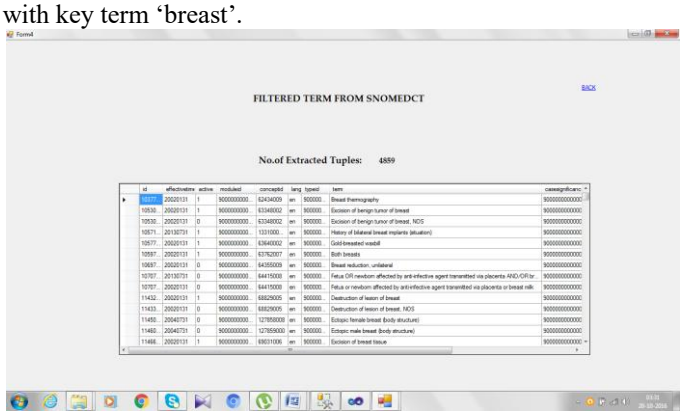


Fig. 3 Query-based SNOMED Subset extraction on key term 'breast'

The results of SNOMED subset extraction using key-terms is shown in Fig. 4. Key-term 1 is 'Breast' which is appended to the key terms listed below.

Key_term2	Terms_ extracted	Newterms
Cancer	54	0
Pm	0	0
Stage	27	8
Pathological	0	0
Specimen	60	0
Procedure	446	7
Malignant	118	31
Benign	0	0
Mastectomy	26	20
Carcinoma	145	0
Tumour	74	2
Tumor	184	13
Lymph node	65	40
Metastatic	1	0
pTNM	0	0
pT	135	92
pN	83	75

Fig. 4 SNOMED subset extraction using key-term-based queries

The summary shows that the system successfully extracted 1223 tuples on female breast cancer related concepts from SNOMED. The top 3 key-terms that yielded the highest number of hits are Breast + procedure, Breast + tumor and Breast + carcinoma. Most of the works done earlier relating to SNOMED annotation were performed offline, while the CDSS, annotates reports at the time of generation of a report. The SNOMED subset was successfully used to annotate reports and cancer staging was performed.

In a CDSS, the main components are the Knowledgebase, The Rule Engine and the Graphical User Interface (GUI). The knowledgebase has the dataset, AJCC protocols, CAP checklist and the SNOMED database and the SNOMED subset, which are used by the Rule Engine to perform the Pathological classification and cancer staging processes. The results of all processes are presented to the user through the GUI. While Pathological classification and Cancer staging are processes that are patient-centric and contribute to better patient-care, the SNOMED annotation process is used in order to standardize the reports with a global terminology standard and check if it would improve the performance of the pathological and cancer staging processes. According to the hypothesis, we do not expect the system's performance to be enhanced through SNOMED annotation and subsequent Pathological processes. SNOMED annotation works done earlier have inferred that the Annotation of a medical report with standard terms and their Ids ensure consistency in language, and structuring of the textual content, thereby improving efficiency in machine-readability, retrieval and analysis. The annotation mechanism have also resulted in an improved extraction process. Another benefit of annotation of reports using standard terminology is that it would enable the pathology report to be transmitted without loss of meaning across organizational and demographic boundaries.

Annotation of Pathology reports is performed in two modes - Annotation on a group of reports in offline batch processing mode on archived reports and Annotation on individual reports online, at the time of generation of the report. In the first annotation method, the medical terms in the generated Lexicon are read and if a matching term is found in the report, the term is tagged. The tags used are <t> indicating the beginning and </t> indicating the end of the term. The tagged terms are annotated by the system with its SNOMED term and its corresponding Term Id. For example, if a term *ductal carcinoma in situ* is read from the lexicon and found in the text, it is tagged as <t> *ductal carcinoma in situ* </t> and replaced by its exact SNOMED term and its SNOMED code '86616005'. The Pathologist might have reported the '*ductal carcinoma in situ'* as 'duct. carc.' or 'DC in situ' or in any other short form, since reporting is done using natural language text. Replacing the term that is written by the Pathologist by the SNOMED term standardizes the report. The SNOMED annotation on an archived report is shown in Fig. 5.

Fig. 5 SNOMED-Annotated Pathology report

In the second annotation method, when a user types the pathology report, an *order-independent auto completion algorithm* lists a set of SNOMED terms that match the prefixes of the words typed, from the SNOMED subset. For example, if the user types 'right modified radical mastectomy' or 'ri mod rad mast' or 'rig rad mas mod', the system would display 'right modified radical mastectomy' as one of the options. The user can select the term. Upon selection, the term in the report is replaced by the term found in the SNOMED database and its Term-Id. In the second method, the annotation by the system occurs every time a medical term is typed by the user. Fig. 6 shows the screenshot of auto-annotation by the CDSS, at the time of generation of the report by the Pathologist.

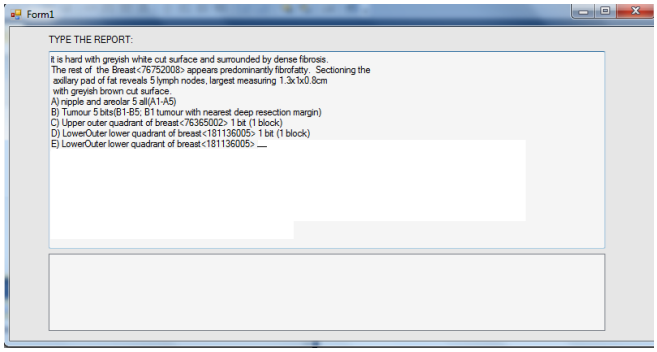


Fig. 6 Online SNOMED-Annotation of a Pathology report

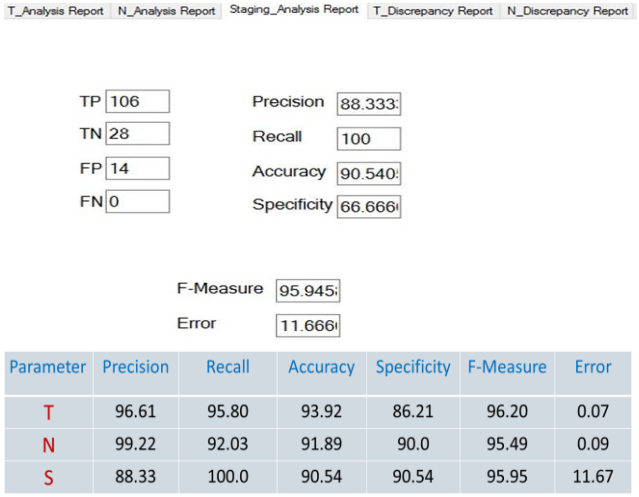
Both archived and newly generated reports can be annotated by the system. The main focus of the CDSS is to perform the Pathological processes on the natural language reports and the SNOMED annotated reports. The AJCC protocol, 7TH Edition was used for Pathological classification and cancer staging. Pathological Classification is named pTNM, where the first letter is the TNM descriptor. In a TNM descriptor, *p* represents pathological, *m* represents multiple foci of invasive carcinoma, *r* represents recurrent, and *y* represents post treatment). T represents Tumour, N represents Lymph node and M represents Distant Metastasis. pT can take the values pTX, pT0, pTis (DCIS), pTis (Paget), pT1, pT1mi, pT1a, pT1b, pT1c, pT2, pT3, pT4 (pT4a, pT4b, pT4c, pT4d). pN can take the values pNX, pN0, pN0 (i-), pN0 (mol+), pN1mi, pN1a, pN2a, pN3a. pM can take the values pMX, pM0, pM1.

A staging system for cancer indicates how far the cancer has spread. There are two methods of staging, namely Clinical staging and Pathological staging and out of the two, Pathological staging is more accurate. The CDSS derives the Pathological classification and cancer staging from the Impression or Summary section of the report. The Rule Engine has the AJCC protocols described using the *Event-Condition-Action Model*. In this model, each rule is read as "*ON event IF condition THEN action*". The CDSS triggers the rules using both the pathology reports written in natural language and those annotated using SNOMED CT. The pattern matching rules for Pathological classifications are written using numerical and non-numerical conditions to extract the relevant details from the reports and triggered by the rules of the CDSS to determine the pTNM. There are 56 rules for cancer staging based on AJCC guidelines for breast cancer staging. The results of the system's performance on plain text and annotated text were compared and analyzed to check the hypothesis.

IV. RESULTS AND ANALYSIS

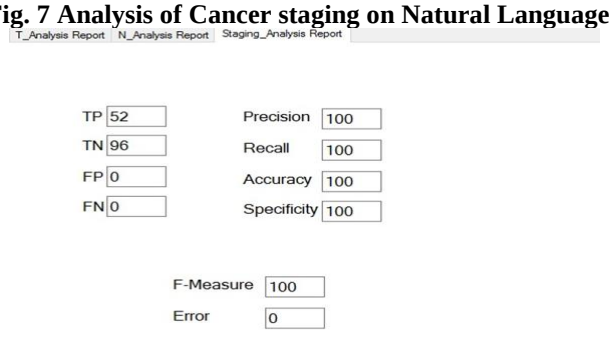
The Pathological classification and cancer stage were determined for the patients by the rule engine of the CDSS and the results were analyzed by finding the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values. The Gold standard for the evaluation was set by the Pathologists of the hospital through a manual process. The evaluation parameters used in the analysis are Precision, Recall, Specificity, Accuracy, F-measure and Error Rate.

Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset.
The best error rate is 0.0, whereas the worst is 1.0. *Accuracy* (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0. It can also be calculated by 1 – ERR. *Sensitivity* (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called *recall* (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0. *Specificity* (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called *true negative rate* (TNR). The best specificity is 1.0, whereas the worst is 0.0. *Precision* (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is also called *positive predictive value* (PPV). The best precision is 1.0, whereas the worst is 0.0. *False positive rate* (FPR) is calculated as the number of incorrect positive predictions divided by the total number of negatives. The best false positive rate is 0.0 whereas the worst is 1.0. It can also be calculated as 1 – specificity. *F-score* is a harmonic mean of precision and recall. Fig. 6 and 7 present the CDSS output for cancer staging analysis. In cancer staging M is assumed to be M0 as it is clinically determined parameter. Fig. 7 and Fig. 8 present the results of analysis of cancer staging performed on natural language text and SNOMED annotated text.



Parameter	Precision	Recall	Accuracy	Specificity	F-Measure	Error
T	96.61	95.80	93.92	86.21	96.20	0.07
N	99.22	92.03	91.89	90.0	95.49	0.09
S	88.33	100.0	90.54	90.54	95.95	11.67

Fig. 7 Analysis of Cancer staging on Natural Language text



Precision	Recall	Accuracy	Specificity	F-Measure	Error	Parameter
93.98	100.0	96.6	92.88	96.89	0.06	T
92.45	100.0	97.30	95.96	96.08	0.08	N
100.0	100.0	100.0	100.0	100	0	S

Fig. 8 Analysis of Cancer Staging on SNOMED-annotated text

The results disprove the hypothesis and indicate that SNOMED annotated reports gave accurate information extraction of Pathological parameters, resulting in improved results in Pathological classification and cancer staging processes. We thus infer that SNOMED annotation improves the performance of the CDSS is providing support for Pathological decisions in breast cancer domain.

V. CONCLUSION

The objective of the work was to develop a CDSS prototype to perform Pathological classification and cancer staging on breast cancer patient records. The SNOMED subset for the breast cancer pathology domain was extracted using two methods and annotation of the reports were done in offline and online modes. The important processes in the CDSS that support decision making by the Pathologist are the Pathological classification and cancer staging. These processes were performed both on natural language textual reports and SNOMED annotated reports. The use of standard AJCC protocol for cancer staging and globally accepted medical vocabulary such as SNOMED yielded better results in the staging process. The accuracy of automated systems in medical domain, especially in a task as critical as cancer staging is of vital importance, as it involves diagnostic and treatment decisions for the patient. This critical factor necessitates that reports be annotated and processed for better results. Annotation of the reports using SNOMED also makes it possible to analyze and understand the patient population through application of queries.
The work clearly indicates that between cancer staging process on natural language text and the SNOMED annotated text, the process on annotated text yields best results. As extension of this work, a similar CDSS process can be performed on reports of other cancer domains for improved decision making. This can be done by incorporating rules to the CDSS for the other cancer domains using standard medical procedures and protocols.

ACKNOWLEDGMENT

The authors thank the Department of Pathology, Christian Medical College and Hospital, Vellore for providing them with the sample data for their study. Our special thanks to Dr. Joy John Mammen, Dr. Marie Therese Manipadam and Dr. Gunadala Ishitha, Department of Pathology, CMC, Vellore for sharing their domain expertise in the field of Breast Cancer Pathology and evaluations for correctness in medical perspective. The contributions of Ms Sreeja and Mr. Pradeep Vignesh, former students of the Department of Computer Science, Madras Christian College is also deeply appreciated.

REFERENCES

1. World Health Organization. The International classification of Diseases (ICD) [Internet] Geneva, Switzerland: World Health Organization; c2016. http://www.who.int/whosis/icd10.
2. International Health Terminology Standards Development Organisation Copenhagen, Denmark: International Health Terminology Standards Development Organisation; c2016. http://www.ihtsdo.org.
3. Logical Observation Identifiers Names and Codes (LOINC), Indianapolis (IN) The Regenstrief Institute Inc.; c2016, https://loinc.org.
4. http://mohfw.nic.in/documents/electronic-health-record-ehr-standards-india-2016.

5. Fernandes, Ana & Alves, Pedro & Jarman, Ian & Etchells, Terence & Fonseca, Jose & Lisboa, Paulo. (2010). A Clinical Decision Support System for Breast Cancer Patients. 314. 122-129. 10.1007/978-3-642-11628-5_13.
6. Brigitte Seroussi, Jean-Baptiste Lamy, Naiara Muro, Nekane Larburu, Booma Devi Sekar, Gilles Guezennec and Jacques Bouaud, Implementing Guideline-Based, Experience-Based, and Case-Based Approaches to Enrich Decision Support for the Management of Breast Cancer Patients in the DESIREE Project Decision Support Systems and Education J. Mantas et al. (Eds.) © 2018 The authors and IOS Press, doi:10.3233/978-1-61499-921-8-190.
7. Wanyonyi Peter Simeon, A Decision Support System for the Diagnosis of Breast Cancer using Fuzzy Logic and Case Based Reasoning, University of Nairobi, School of Computing and Informatics, July 2016
8. A Novel Knowledge Base Decision Support System Model for Breast Cancer Treatment R. K. Sinha, Dr. M.M. Manohara Pai, Dr. M. S. Vidyasagar, Dr. B.M. Vadhiraja, Sri Lanka Journal of Bio-Medical Informatics 2010;1(2):97-103. DOI: 10.4038/sljbmi.v1i2.1609.
9. Inderpreet Kaur, Scope of Clinical Decision Support System (CDSS) in Healthcare Industry, IIHMR University, https://www.iihmr.edu.in/student-dissertation/2015-2017/scope-of-clinical-decision-support-system-cdss-in-healthcare-industry.
10. Al-Hablani B, The Use of Automated SNOMED CT Clinical Coding in Clinical Decision Support Systems for Preventive Care. Perspect Health Inf Manag. 2017;14(Winter):1f. Published 2017 Jan 1.
11. AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM. Automated Extraction and Classification of Cancer Stage Mentions from Unstructured Text Fields in a Central Cancer Registry. AMIA Jt Summits Transl Sci Proc. 2018;2017:16–25. Published 2018 May 18.
12. Johanna Johnsi Rani G, D. Gladis, M. T. Manipadam and G. Ishitha, "Breast cancer staging using Natural Language Processing," 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, 2015, pp. 1552-1558, doi: 10.1109/ICACCI.2015.7275834.
13. Erik Cambria, Bebo White, Jumping NLP Curves: A Review of Natural Language Processing Research, IEEE Computational intelligence magazine, pp 48-57, May 2014
14. Anthony N Nguyen et al., Symbolic rule-based classification of lung cancer stages from free-text pathology reports, Journal of the American Medical Informatics Association (JAMIA) , 17:440-445, 2010.
15. Nguyen, Moore, Lawley, Hansen, Colquist, Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications, Stud Health echnol Inform. 2011;168:117-24.
16. Napolitano G, Fox C, Middleton R, Connolly D, Pattern-based information extraction from pathology reports for cancer registration, Cancer causes control, 2010 Nov;21(11):1887-94. doi: 10.1007/s10552-010-9616-4. Epub 2010 Jul 23.
17. AJCC: Breast. In: Edge SB, Byrd DR, Compton CC, et al, eds.: AJCC Cancer Staging Manual. 7th ed. New York, NY: Springer, 00 347-76, 2010.
18. Nelson HD, Weerasinghe R, Martel M, Bifulco C, Assur T, Elmore JG, et al. Development of an electronic breast pathology database in a community health system. J Pathol Inform 2014;5:26.
19. Buckley JM, Coopey SB, Sharko J, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. Journal of Pathology Informatics. 2012;3:23. doi:10.4103/2153-3539.97788.
20. Aleksovski Z., Sevenster M. (2011) Identifying Breast Cancer Concepts in SNOMED-CT Using Large Text Corpus. In: Szomszor M., Kostkova P. (eds) Electronic Healthcare. eHealth 2010. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 69. Springer, Berlin, Heidelberg
21. Krystyna Milian, Zharko Aleksovski, Richard Vdovjak, Annette ten Teije, Frank van Harmelen, Identifying Disease-Centric Subdomains in Very Large Medical Ontologies: A Case-Study on Breast Cancer Concepts in SNOMED CT. Or: Finding 2500 Out of 300.000, AIME 2009 Workshop KR4HC 2009, Verona, Italy, , Revised Selected and Invited Papers, pp 50-63, 2010, DOI 10.1007/978-3-642-11808-1_5, July 19, 2009.
22. Carlos Rodriguez-Solano, Leonardo Lezcano, Miguel-Angel Sicilia, Automated Generation of SNOMED CT Subsets from Clinical Guidelines, Chapter 13, Information Systems and Technologies for enhancing Health and Social care, Medical Information Reference, Copyright, IGI Global, .DOI: 10.4018/978-1-4666-3667-5.ch013, 2013.
23. Ahmed M. Alaa, Kyeong H. Moon, William Hsu, Mihaela van der Schaar, Fellow, IEEE, ConfidentCare: A Clinical Decision Support System for Personalized Breast cancer Screening, https://arxiv.org/abs/1602.00374v1 [cs.LG], 2016
24. Spencer Robinson , Veronique Poirier, Sam Watson, Using Cancer Decision Support Tools to support the early diagnosis of cancer, Accelerate, Coordinate, Evaluate (ACE) Programme, An early diagnosis of cancer initiative supported by NHS England, Cancer Research UK and Macmillan Cancer Support
25. Ronak Sumbaly, N. Vishnusri. S. Jeyalatha, Diagnosis of Breast Cancer using Decision Tree Data Mining Technique, International Journal of Computer Applications (0975 – 8887), Volume 98– No.10, July 2014.
26. Henry NL, Bedard PL, DeMichele A, Standard and Genomic Tools for Decision Support in Breast Cancer Treatment, American Society of Clinical Oncology Educational Book. 2017;37:106-115, PMID:28561710, DOI:10.14694/EDBK_175617
27. Dennis Lee, Ronald Cornet, Francis Lau, Nicolette de Keizer, A survey of SNOMED CT implementations, Journal of Biomedical Informatics 46, pp. 87–96, 2013.
28. SNOMED Clinical Terms, http://www.ihtsdo.org/snomed-ct.
29. Lin C-H, Wu N-Y, Lai W-S, Liou D-M. Comparison of a semi-automatic annotation tool and a natural language processing application for the generation of clinical statement entries. Journal of the American Medical Informatics Association : JAMIA. 2015;22(1):132-142. doi:10.1136/amiajnl-2014-002991.

AUTHORS PROFILE

Mrs. G. Johanna Johnsi Rani is currently working as Assistant Professor in the Department of Computer Science, Madras Christian College (Autonomous) in Chennai. She has 31 years of Teaching experience and 5 years of Research experience. She is pursuing her research in Computer Science in the field of Natural Language Processing. She has 8 publications in International Journals and has presented research papers in 12 International Conferences.

Dr. (Mrs.) Gladis Christopher is the Principal of Bharathi Women's College in Chennai since 2018. Earlier she headed the PG and Research Department of Computer Science, Presidency College, Chennai. She has guided more than fifteen M.Phil and Ph.D Research scholars and has been an active researcher in the areas of Neural Networks, Image Processing, and Data Mining.

Dr. Joy John Mammen is currently the Professor and Head of the Department of Transfusion Medicine & Immunohaematology in CMC. His main research work focuses on Laboratory and Healthcare Informatics, Digital Imaging in Pathology (Acquisition, Management and Analysis), Error reduction & Tracking (Barcoding solutions for the clinical lab), Lab Decision Support (Algorithms for automated cell counter area), Patient care (Nosocomial infection - active surveillance) and National External Quality Assessment