Automated SNOMED CT Mapping of Clinical Discharge Summary Data for Cardiology Queries in Clinical Facilities

Abdul Aziz Latip[1], Ma. Stella Tabora Domingo[1], Ismat Mohd Sulaiman[2], Tengku Nurulhuda Tengku Abd Rahim[1]
[1]Artificial Intelligence Lab, MIMOS Berhad, Technology Park Malaysia
[2]Health Informatics Centre, Planning Division, Ministry of Health Malaysia
Email: {abdaziz.latip, stella.domingo, huda.rahim}@mimos.my, drismat@moh.gov.my

***Abstract*—Heart disease has remained the leading cause of death among Malaysians for 13 years from 2005 to 2017. As it has become the prominent factor of death in Malaysia, the intention is to improve the accuracy of query for cardiology related cases as it is the primary source of analytical data for heart disease. Choosing the right terminology is one of the criteria to improve the accuracy as the clinical term can be mapped as much as possible. Therefore, Systematized Nomenclature of Medicine Clinical Term (SNOMED CT) has been selected for implementation as it is known as the most comprehensive, multilingual clinical healthcare terminology in the world. This paper presents the implementation to enrich and increase the result accuracy by automatically mapping the Clinical Discharge Summary using several techniques in Natural Language Processing (NLP) with SNOMED CT. By observing the trend and pattern of data, a facility or ministry can plan one step ahead, through prevention or future planning. Therefore, the accuracy of the result is the key factor to derive the outcome.**

***Index Terms*—cardiology, terminology, SNOMED CT, NLP**

I. Introduction

Cardiology is a branch of medicine that deals with the disorders of the heart as well as some parts of the circulatory system. The field includes medical diagnosis and treatment of congenital heart defects, coronary artery disease, heart failure, etc. [1]. Clinical Discharge Summary Data are information that is recorded via Hospital Information System (HIS) based on Clinical Discharge Summary Notes prepared by a Doctor in free text.

Cardiology Queries refers to the query of any cardiology-related cases based on cumulative data from either one or a group of facilities in Malaysia; usually to observe following results: (i) Total number of patients admitted for a particular disease; (ii) Total number of patient deaths for a particular disease, (iii) Total number of patient visits for a particular disease, (iv) Total number of patients who have undergone a certain procedure based on selective disease. Inaccuracies exist during the query of any cardiology-related case because the source data is written in free text, which will be explained further at the problem statement section.

The primary purpose of SNOMED CT is to encode the meanings that are used in health information and to support the effective clinical recording of data with the aim of improving patient care.

The implementation described in this paper will automatically map the clinical terms found in the source data from Group of Malaysian Clinical Facilities Database to the SNOMED CT to support Cardiology Queries. Those data were originally from Clinical Discharge Summary Reports written in free text that were recorded into a database using Hospital Information System (HIS). The exported data contains the following sections: (i) patient information; (ii) ward information; (iii) reason for admission; (iv) historical related diseases; (v) lab findings; (vi) treatment; (vii) patient deceased status, (viii) diagnoses; (ix) medical examiners comments; (x) procedure; (xi) discussion; (xii) recommendations. All information is entered as free text except for: (i) personal data; (ii) ward information; (iii) patient deceased status;

In summary, this paper discusses our approach for automated mapping in Clinical Discharge Summary Data using SNOMED CT. The paper is organized as follows: Section 2 will elaborate related work. Section 3 describes the problem statement, Section 4details out methodology that has been applied. Section 5 describes the findings and discovery on the implemented process. Section 6 shows the results obtained in current implementation. Section 7 is conclusion.

II. Related Work

The application of natural language processing methods to clinical free-text is of growing interest for both health care practitioners and researchers. For our exploration, we aim to provide analytic result for cardiology queries cases based on Clinical Discharge Notes from Malaysian Hospitals and Clinics. We found that the information from Clinical Discharge Notes were written as free text and considered as unstructured data.

Manuscript received August 17, 2020; revised October 24, 2020.

doi: 10.18178/ijpmbs.10.1.8-16

Currently,cardiology query is generated by using structured data obtained via "Sistem Maklumat Rawatan Pesakit" (SMRP) or Patient Treatment Information System by using ICD-10 as references. This information is recorded after Clinical Discharge Information is entered in the HIS. We have noticed that, it requires a longer time to get the results of cardiology-related query and requires major effort to produce results based on Subsumption and negative statement filtered using existing system.

As our reference, firstly, we looked at the paper "Automatic Matching of ICD-10 codes to Diagnoses in Discharge Letters". This is because we wanted to know how the possible method of automatic matching works with Discharge Letters. The objective of that paper is to automatically map discharge letters using ICD-10 for Bulgarian Hospitals. Since there is no international format for discharge letters, they prepared the standard template for Bulgarian Hospital based on mandatory structure, which is published in the Official State Gazette within the legal Agreement between the Bulgarian Medical Association and the National Health Insurance Fund [2]. This is standardized the input for automatic matching. The method is using SVM Classifier for automatic assignment of ICD-10 code. For the preprocessing part, several activities have been done including splitting, tokenization, diagnoses extraction, abbreviation expansion, transliteration and Latin terminology processing. For our research, we did a similar approach by using the prepared standard template for Malaysian Hospital and Clinic for Cardiology Related by identifying the common fields that are required for matching. For the preprocessing part, we have a different approach and tokenization is a common technique during preprocessing.

Second, we looked into the method from the paper "Exploiting Fast Classification of SNOMED CT for Query and Integration of Health Data". We observed the implementation uses a polynomial algorithm applied in Java that is able to classify using SNOMED CT in under 1 minute. The method that has been applied uses identified relationship with SNOMED to provide results based on synonym and subsumption which is similar to our system. In addition, that paper also introduce Local Extension that construct an ontology from an existing terminology and continued by using augmented table for target data, the SNOMED CT expression is appended to additional column that will be used for queries. In that paper, they mentioned OWL as the standard format to represent ontologies but this is not practical for large ontologies like SNOMED CT. For our research, there are two common techniques that are used to retrieve query results based on synonym and subsumption relationship in SNOMED CT. However, we added a negation technique as we found out that there are negative statements in the discharge clinical notes that should be excluded from query results.

III. Problem Statement

A. Synonyms

Synonyms refer to the same concept that can be expressed in different ways. Unlike fully specified names, synonyms are not necessarily unique as the same term may be used to describe more than one concept. In any given language context, a concept may have any number of synonyms that are acceptable for use and must have one synonym that is preferred for use (the preferred term). The synonyms that are preferred or acceptable are specified by a language reference set for the relevant language context [3].

An example in the Clinical Discharge Summary Notes, the clinical term *Insertion of permanent cardiac pacemaker pulse generator and electrode* defined in SNOMED CT could be written in simplified term as *Permanent pacemaker insertion* or *Permanent pacemaker* or *Pacemaker implantation*. Another example is *Acute non-ST segment elevation myocardial infarction* defined in SNOMED CT may be written as the abbreviation NSTEMI. Table I shows a list of fully specified term in SNOMED CT with synonym written in Clinical Discharge Summary Notes. Most of the clinical terms are either written as abbreviation, simplified term or local term but identical with fully specified term defined in SNOMED CT.

TABLE I. Example of Synonym Written in Discharge Clinical Summary Notes

SNOMED CT Fully Specified Term	Type	Similar Term (Synonym)
Insertion of permanent cardiac pacemaker pulse generator and electrode	Procedure	Permanent pacemaker insertion
		Permanent Pacemaker
		Pacemaker implantation
Percutaneous coronary intervention	Procedure	PCI
Coronary artery bypass graft	Procedure	CABG
Ischemic heart disease	Disorder	IHD
		Ischaemic Heart Disease
Acute non-ST segment elevation myocardial infarction	Disorder	NSTEMI

B. Formatted Text

Formatted Text is data that have been formatted in application format by HIS which is recorded and stored in the database. The intention is to display in application in formatted way during information retrieval. As a result the clinical term intended to map with SNOMED CT is mixed up with formatted tag. Fig. 1 shows the term *Sepsis* has been mixed together with tag *</FONT>* along with the other formatting and joined altogether with term *Reticulocyte* becoming one word, whereas these terms are fully specified in SNOMED CT as *Sepsis* and *Reticulocyte* respectively. In this case, the term itself has been joined together with formatting tag as a single word. This is a common case after formatting, wherein one or more clinical terms are mixed together with formatting tag as one word. As a result, these clinical terms may

often be missed out during mapping, leading to inaccurate result when querying cardiology cases.

```
<ADDRESS><STRONG><U><FONT size=3>
DIAGNOSIS:</FONT></U></STRONG></ADDRESS
> <ADDRESS><STRONG><FONT size=3>1)
Presumed
Sepsis</FONT></STRONG></ADDRESS><BR>Coo
mb's negative<BR>Reticulocyte:
3.94%<BR><BR>Diagnosis: <BR>1. presumed
sepsis infant of mother with premature
rupture of membrane &gt;12H, adequately
treated but leucocytosis<BR>2. Neonatal
Jaundice secondary to ABO
incompatibility <BR>MBG:O positive and
BBG: B positive <BR>Coombâs
test:negative<BR><BR>
```

Figure 1. Example of formatted text clinical discharge summary data.

C. Subsumption

Subsumption occurs when one clinical meaning is a subtype of another clinical meaning. If clinical meaning X is a subtype of clinical meaning Y, then Y is said to 'subsume' X and X is 'subsumed by' Y [4]. Fig. 2 shows example of subsumption for Ischemic Heart Diseases. In this diagram we only show four (4) clinical terms, which is subset for Ischemic Heart Disease: (i) Angina; (ii) Chronic Ischemic Heart Disease; (iii) Myocardial Infarction; AND (iv) Acute Ischemic Heart Disease. In addition, there are also clinical term subsets for each of them. In this diagram, we show only up to two clinical term subset for them.

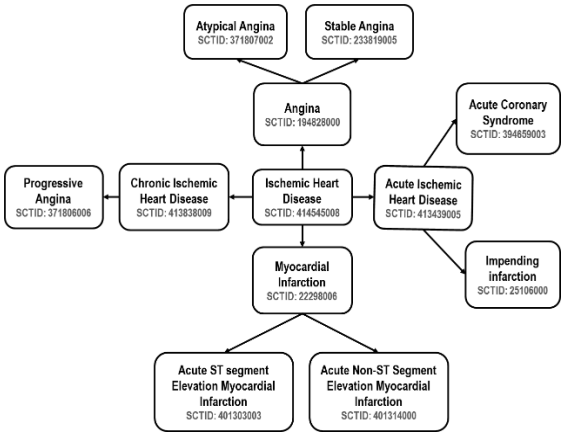

Figure 2. Subsumption for ischemic heart disease in SNOMED CT.

For example the term *Atypical Angina* in cardiology cases is a subtype of *Angina* and the term *Angina* itself is a subset of Ischemic *Heart Disease.* In this case, when query for Ischemic Heart Disease is issued, all the corresponding child terms must be included as a result; not just Ischemic Heart Disease. The following inheritance principle is taken for granted in work on ontologies and terminologies [4]:

- If A is a child of B then all properties of B are also properties of A.
- Additionally, one inheritance principle based on Bernauer's approach to subsumption can be expressed as follows [5]:
- All roles of a parent class must either be inherited by each child or refined in the child.
- This principle can also be formulated from the perspective of the child as follows: Differentia from child to parent should uniquely result in every case either from refinement of the value of a common role or introduction of a new role.

D. Negation

Negation occurs when there are statement claims in the Discharge Clinical Notes denying any clinical term that was not present in the patient. In this case, when query is issued the result must be excluded. Table II shows an example of the variations of negative statements in Discharge Clinical Notes that must be excluded from the result of query. There is a variation of negation words used to deny clinical term present in current diagnosis such as "History of", "No evidence", "not a candidate", "Not suitable", "Absent" and etc.

TABLE II. EXAMPLE OF NEGATIVE STATEMENT IN DIAGNOSIS CONTAIN IN CLINICAL DISCHARGE NOTES

Visit ID	Diagnosis
IP0278135	History of Extensive Anterolateral STEMI in February 2017. - Successfully thrombolysed.
IP0274732	Coronary artery disease. PCI to LCx. LVEF 62% with History of Inferoposterior STEMI.
IP0275835	no evidence of coronary artery disease.
IP0279237	Triple Vessel Disease (not a candidate for Coronary Artery Bypass Graft)
IP0269076	Coronary artery disease (History of STEMI on 9/1/2017)
IP0265502	Mild coronary artery disease - History of acute lateral STEMI on 9/2/2017 (give STK at Hospital Melaka)
IP0275884	Ischemic Heart Disease Triple Vessel Disease - not suitable for Coronary Angiogram- not keen for repeat angiogram
IP0275962	Tetralogy of Fallot with absent pulmonary valve syndrome

IV. METHODOLOGY

A. SFTP Server

The first step is preparing the SFTP server and creating user ids for each Hospital representative shown in Fig. 3. This allows them to periodically upload the files in JSON Format which contains the extracted discharge summary data from their database into the server. JSON is a suitable format when multiple facilities are uploading data, due to interoperability and lesser data storage. To minimize the error during processing of data into the database, a JSON standard. template has been introduced so that every facility will upload according to the template. The template is designed for querying cardiology related cases, which contains the following information with the same key as identifier: *(i)* patient information; *(ii)* ward information; (*iii*) reason for admission; (*iv*) historical related diseases; *(v)* lab findings; *(vi)* treatment; (*vii*) patient deceased status, *(viii)*

diagnoses; *(ix)* medical examiners comments; *(x)* procedure; *(xi)* discussion; *(xii)* recommendations.

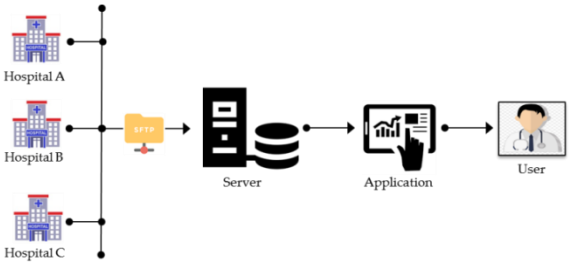

Figure 3. Applying SFTP server in infrastructure architecture.

B. System Architecture

Fig. 4 shows the system architecture, where we break up the entire process into three modules, which are, *(i)* Data Preprocessing *(ii)* Codification; (*iii*) Query. Data preprocessing will remove unnecessary text in the clinical discharge notes. The Codification module tags clinical terms present in Discharge Clinical and will untag if the statement is negative statement. Query is the module that returns the semantic result based on cardiology related queries issued by user.

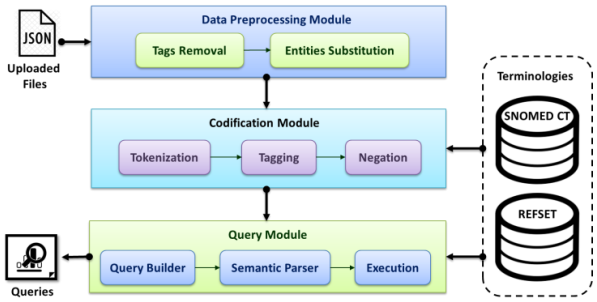

Figure 4. System architecture.

C. Data Preprocessing

In the data preprocessing module, the data extracted from files and uploaded in SFTP is converted into JSON object. To simplify the codification process later, HTML tags inside the data is removed. This is to optimize the time taken for traversing word by word at the codification process, as well as entity searching. By removing unnecessary formatting, the codification process will be more efficient and the time taken will drastically be reduced especially when multiple files are being uploaded. In this process, there are two components; tags removal and entity converter.

Tags removal is where most of the HTML opening and closing tag is completely removed while the embedded content is preserved. In this process, all HMTL tags are removed except the <BR> tag as shown in Fig. 5 which is required to be convert into new line.
 tag is excluded in this process because of two reasons. The main reason is to avoid any word including clinical term before <BR> being joined together with word after to become one word and lead to the possibility of missing terms when matching the clinical term with SNOMED CT. This process is subsequently followed by the Entity Converter. At this level, all remaining HTML characters are converted into UTF8 characters as well as
 Tags. Table III is the list of common HTML entities that undergo the conversion process.

Before Tags Removal

```
<ADDRESS><STRONG><U><FONT size=3>
DIAGNOSIS:</FONT></U></STRONG></ADDRESS>
<ADDRESS><STRONG><FONT size=3>1) Presumed
Sepsis</FONT></STRONG></ADDRESS><BR>Coomb
's negative<BR>Reticulocyte:
3.94%<BR><BR>Diagnosis: <BR>1. presumed
sepsis infant of mother with premature
rupture of membrane &gt;12H, adequately
treated but leucocytosis<BR>2. Neonatal
Jaundice secondary to ABO incompatibility
<BR>MBG:O positive and BBG: B positive
<BR>Coomb's test:negative<BR><BR>History
of presenting illness<BR>Baby referred
from labour room for presumed sepsis,
infant of mother with premature rupture
of membrane &gt;12H, adequately treated.
<BR>Baby was born vigorous with good
APGAR score. Suction: clear<BR>On
arrival, not tachypneic <BR>No signs or
symptoms of respiratory distress<BR>DXT
on arrival: 3.6mmol/L<BR>BP:
76/52mmHG<BR>PR:138<BR>SPO2:100%
```

After Tags Removal

```
DIAGNOSIS: 1) Presumed Sepsis<BR>Coomb's
negative<BR>Reticulocyte:
3.94%<BR><BR>Diagnosis: <BR>1. presumed
sepsis infant of mother with premature
rupture of membrane &gt;12H, adequately
treated but leucocytosis<BR>2. Neonatal
Jaundice secondary to ABO incompatibility
<BR>MBG:O positive and BBG: B positive
<BR>Coomb's test:negative<BR><BR>History
of presenting illness<BR>Baby referred
from labour room for presumed sepsis,
infant of mother with premature rupture
of membrane &gt;12H, adequately
treated.<BR>Baby was born vigorous with
good APGAR score. Suction: clear<BR>On
arrival, not tachypneic <BR>No signs or
symptoms of respiratory distress<BR>DXT
on arrival: 3.6mmol/L<BR>BP:
76/52mmHG<BR>PR:138<BR>SPO2:100%
```

After Entity Converter

```
DIAGNOSIS: 1) Presumed Sepsis
Coomb's negative
Reticulocyte: 3.94%

Diagnosis:
1. presumed sepsis infant of mother with
premature rupture of membrane > 12H,
adequately treated but leukocytosis
2. Neonatal Jaundice secondary to ABO
```

```
incompatibility
MBG:O positive and BBG: B positive
Coomb's test:negative

History of presenting illness<BR>Baby
referred from labour > room for presumed
sepsis, infant of mother with premature
rupture of membrane 12H, adequately
treated. Baby was born vigorous with good
APGAR score. Suction: clear
On arrival, not tachypneic
No signs or symptoms of respiratory
distress
DXT on arrival: 3.6mmol/L
BP: 76/52mmHGD.
PR:138
SPO2:100%
```

Figure 5. Text preformatting result.

TABLE III. COMMON HTML ENTITIES THAT UNDERGO CONVERSION PROCESS

HTML Character	UTF8	Description
&gt;	>	Greater than
&lt;	<	Less than
		Non-breaking space
 		New Line

D. Codification

Codification is the module, where clinical terms in the data that has been preprocessed is tagged with SNOMED CT concepts. This process is divided into three components *(i)* Tokenization *(ii)* Negation; (*iii*) Tagging. To optimize the codification process, only unstructured data that contain clinical terms will be selected for codification as shown in Table IV. Thus, patient information, ward information, patient deceased status is excluded.

TABLE IV. LIST OF SECTION IN CODIFICATION PROCESS

Section	Type	Codified
Patient Information	Structured	No
Ward Information	Structured	No
Patient Deceased Status	Structured	No
Reason For Admission	Unstructured	Yes
Historical Related Diseases	Unstructured	Yes
Lab Findings	Unstructured	Yes
Treatment	Unstructured	Yes
Diagnoses	Unstructured	Yes
Medical Examiners Comments	Unstructured	Yes
Procedure	Unstructured	Yes
Discussion	Unstructured	Yes
Recommendations	Unstructured	Yes

In the tokenization process, the entire text in each selected section is broken up into its individual word for subsequent processing. Tagging is the process where each word that was broken up as a result of the tokenization process is grouped into a phrase and then matched to clinical terminologies.

Fig. 6 shows raw text that has been broken up into its individual word is grouped with one another either before or after into a phrase. Subsequently, tagging will be applied if it is a match with a clinical term defined in the terminology. In this case, the word coronary, artery and disease match with Coronary Artery Disease defined in SNOMED CT was tagged.

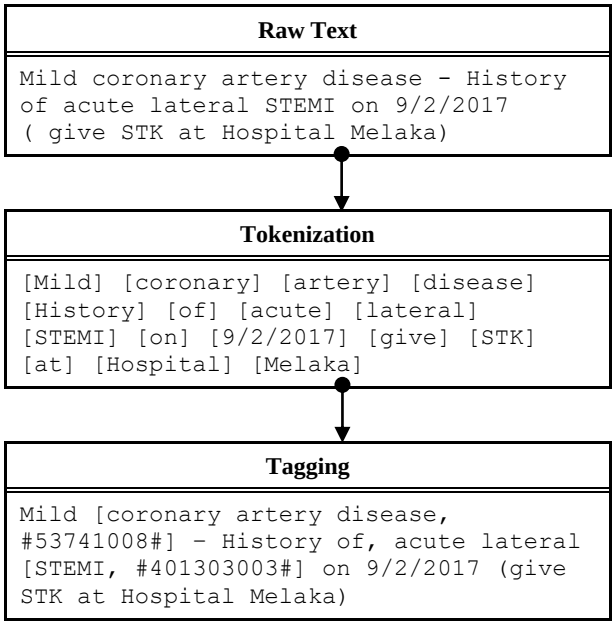

Figure 6. Tagging process.

Negation is the process of detecting negating word in the statement after the tagging process. Generally, the process converts affirmative to negative statement by untagging the clinical terms if the negating word exists in the statement such as "history of", "absent", "to rule out" and etc. In addition, information is added at the end of statements as evidence of negation. The purpose of negation is to exclude the negative clinical term from queries later. Fig. 7 shows tagged text containing the word "history" that denies the next predicate. In this process, the term that was previously tagged has been revoked and evidence of negation was appended for references.

E. Query

Query is the module that returns the semantic result based on cardiology related queries issued by the user. The result will include synonym and subsumption, whereas negation is excluded as discussed in the problem statement. The module is divided into 3 processes, which are: Raw Query, Synonym Query and Subsumption Query. Fig. 8 shows that the user intended to query Ischemic Heart Disease Cases in Hospital A. In Raw Query Transformation, case inquiry is converted in plain native SQL as raw query. The raw query will return exact

result by keyword in any text case. In this process, synonym and subsumption is not included whereas negative statement is not filtered yet.

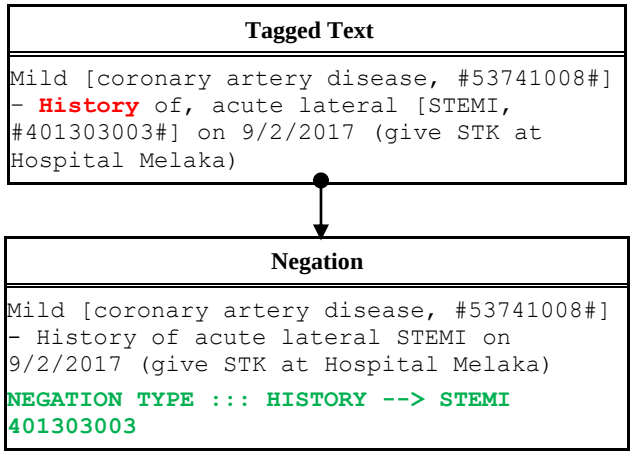

Figure 7. Negation process.

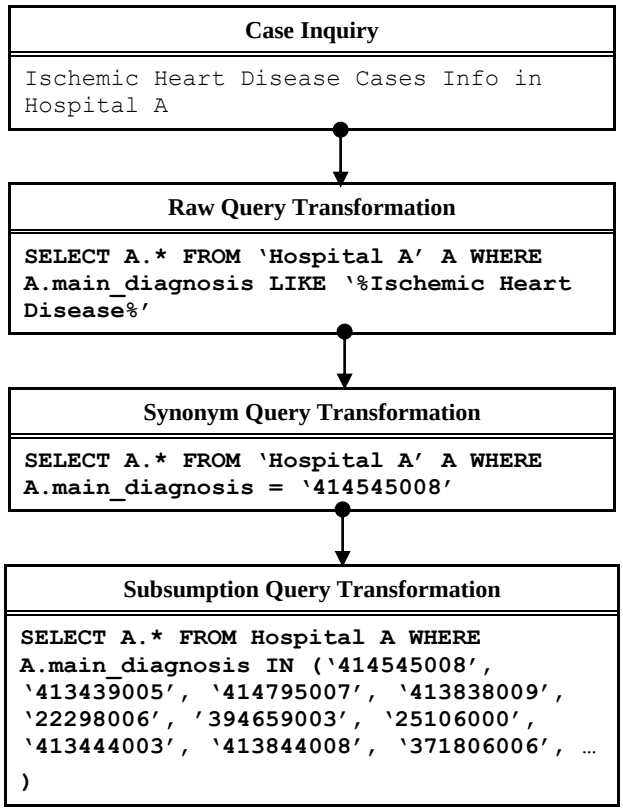

Figure 8. Queries process flow.

Table V shows the result returned based on raw query using sample data. Based on sample data, there are 4 results including 1 record that contains negative statement.

Synonym Query Transformation is the process of transforming raw query into synonym query, where the search keyword is converted into concept id. Generally the process will return the results by concept level without negative statement. The results returned are based on the clinical term that has been tagged in the tagging process. Negative statements are excluded as the clinical term was untagged during the negation process at the Codification module. Based on case inquiry in Figure 8, the result should return the record that contains IHD and Ischaemic Heart Disease as it carries the same concept id with Ischemic Heart Disease.

TABLE V. SAMPLE RESULT ISCHEMIC HEART DISEASE CASES INFO IN HOSPITAL A FOR SYNONYM QUERY

Visit ID	Diagnosis
IP0263950	[**Ischemic heart disease**, #414545008#]. [NSTEMI, #401314000#], [PCI, #415070008#] to [LAD, #29562000#].
IP0268630	[**Ischemic Heart Disease**, #414545008#] [Percutaneous Coronary Intervention, #415070008#] to [Left Main Stem, #3227004#] / [Left Anterior Descending, #59438005#] (2014)
IP0272610	To rule out **Ischemic Heart Disease** **NEGATION** TYPE :::NEGATED-->**Ischemic Heart Disease** 414545008
IP0280608	Mild [**ischemic heart** disease, #414545008#]
IP0278881	[**Ischaemic Heart Disease**, #414545008#] ([Triple Vessel Disease, #233817007#])
IP0252058	[**IHD**, #414545008#] - [PCI, #415070008#] - [RCA, #13647002#] – BVS
IP0267862	[**IHD**, #414545008#] - [LAD, #29562000#] [PCI, #415070008#] 2008
IP0268438	[**Ischaemic heart disease**, #414545008#] Two vessel disease CTO IDCMP EF 30% VT
IP0268321	[**Ischaemic heart disease**, #414545008#] - [Percutaneous coronary intervention, #415070008#] to [Left anterior descending, #59438005#]
IP0263223	To Rule Out **Ischaemic Heart Disease** For Aortic Valve Replacement (Redo) on April 2017 **NEGATION** TYPE :::NEGATED-->**Ischaemic Heart Disease** 414545008 NEGATION TYPE :::NEGATED-->Aortic Valve 181287002

Subsumption Query Transformation is the process of transforming the synonym query into a subsumption query. In general, the process will return all the clinical terms that has been identified in the synonym query along with all the descendants in the terminology. Fig. 9 shows all descendant clinical term for Ischemic Heart Disease.

Table VI shows sample result for the query of Angina and its descendants. The result from negative statement has been filtered out as those clinical terms are untagged during negation process earlier. Since the searching is based on concept id, therefore the entire synonym for descendant is also included. As a result, clinical term like "UA" **(SCTID**: **4557003)** and "Angina Pectoris" **(SCTID**: **194828000)** are included since they are synonyms for Unstable Angina **(SCTID**: **4557003)** and Angina **(SCTID**: **194828000)** respectively as the concepts id are identical. Based on sample data, 10 records were found for subsumption result on Angina.

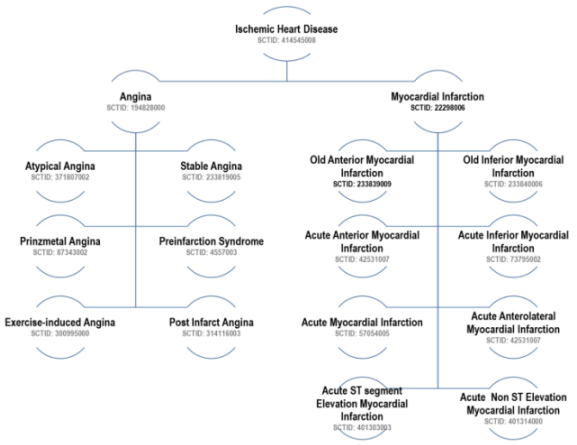

Figure 9. Subsumption relational diagram for ischemic heart disease.

TABLE VI. SAMPLE SUBSUMPTION RESULT OF ANGINA

Visit ID	Diagnosis
IP0276085	[Coronary artery disease, #53741008#] - [angina, #194828000#] equivalent (short of breath)
IP0281294	[Angina, #194828000#] induced by anemia
IP0277749	[Angina Pectoris, #194828000#] : [Positive, #10828004#] stress test on January 2017
IP0265198	[Stable Angina, #233819005#] Musculoskeletal pain
IP0267656	[Exertional Angina, #300995000#] [Percutaneous Coronary Intervention, #415070008#] to [Left Anterior Descending, #59438005#] / [Left Circumflex Artery, #362036002#], Plain Old Balloon [Angioplasty, #11101003#] to [Instent Restenosis, #421327009#] [Left Anterior Descending, #59438005#] / [Left Circumflex Artery, #362036002#].
IP0276171	[Atypical angina, #371807002#]
IP0277102	[Post infarct angina, #314116003#] - [Triple Vessel Disease, #233817007#]
IP0277633	[Unstable angina, #4557003#] - Single vessel disease
IP0266040	[UA, #4557003#] - [normal Coronary Angiogram, #168938005#]
IP0268626	[Prinzmetal angina, #87343002#] Kounis Syndrome

Similarly like Angina, Table VII shows sample result for the query of Myocardial Infarction and its descendants. The result for synonym is also included. Therefore "Old Inferior MI" (**SCTID**: **233840006**), "Acute Anterior MI" (**SCTID**: **54329005**) and "NSTEMI" (**SCTID**: **401314000**) are included as their concepts are identical with Old Inferior Myocardial Infarction (**SCTID**: **233840006**), Acute Interior Myocardial Infarction (**SCTID**: **54329005**) and Non ST Elevation Myocardial Infarction (**SCTID**: **401314000**) respectively.

TABLE VII. SAMPLE SUBSUMPTION RESULT OF MYOCARDIAL INFARCTION

Visit ID	Diagnosis
IP0278517	Missed Anteroseptal [Myocardial Infarction, #22298006#]
IP0276713	Anterior [ST Elevation, #164931005#] [Myocardial Infarct, #22298006#] [PCI, #415070008#] to [Proximal, #40415009#] to [mid, #255562008#] [LAD, #29562000#]
IP0274395	[Old anterior myocardial infarction, #233839009#], possible [old inferior myocardial infarction, #233840006#] [Positive, #10828004#] stress echo - [coronary artery disease, #53741008#] ischaemia Impaired LV function
IP0280489	[Old Inferior Myocardial Infarction, #233840006#]
IP0279027	[IHD, #414545008#] [Old inferior MI, #233840006#] ([thrombolysis, #51308000#])
IP0281267	[Acute Myocardial Infarction, #57054005#] Mons pubis abscess Acute at delirium secondary to [Acute Myocardial Infarction, #57054005#]
IP0272010	[Acute Anterior Myocardial infarction, #54329005#] Kilip 1.
IP0277147	[Acute Anterior MI, #54329005#]. ([Failed, #103709008#] [thrombolysis, #51308000#])
IP0264075	[Acute Anterolateral Myocardial Infarction, #70211005#]
IP0282015	[Acute Inferior Myocardial Infarction, #73795002#]
IP0262908	[STEMI, #401303003#] (Extensive Anterolateral) Primary [Percutaneous Coronary Intervention, #415070008#] to [Left Anterior Descending, #59438005#]
IP0272723	Stabilization of [NSTEMI, #401314000#]
IP0265072	[Non ST Elevation Myocardial Infarction, #401314000#]

V. FINDINGS

Based on our results, we found that typing error from discharge summary data causes inaccuracy of the result due to the following consequence: (i) Missed from tagged; (ii) Improper tagging either tagged as subset or parent concept.

Table VIII shows sample data that contain typing errors such as "post infract angina" was written wrongly instead of Post Infarct Angina at the first record. As a consequence, the Angina was tagged instead of Post Infarct Angina. Similarly on the second and third record, the term "A cute inferior STEMI" and "Acute Interior STEMI" was written wrongly instead of Acute Inferior STEMI. Therefore STEMI as a parent concept has been tagged. For the last record, "CoronaryArteryDisease" was not mapped because Coronary, Artery and Disease was joined together as a single word and could not be matched with Coronary Artery Disease. In this case, some results are not returned as the typing errors were not included. In another case, there will be no result or a few records returned when query is issued with the clinical term at the topmost children level. One of the possible

methods to reduce typing error issue is by implementing auto completion or auto suggestion or clinical term highlighting at Hospital Information System. However such implementation may be costly when considering a group of hospitals. In addition, operation at each hospital may differ from the others and is required to be customized at facility level. Other than that, some other facility may unable to implement and will need to replace with a new Hospital Information System.

TABLE VIII. SAMPLE DATA THAT CONTAIN TYPING ERROR

Visit ID	Diagnosis
IP0273331	post infract [angina, #194828000#] Arrhythmias post infract
IP0273703	A cute inferior [STEMI, #401303003#] ([Triple vessel disease, #233817007#] + [Left main stem, #3227004#])
IP0272495	Acute interior [STEMI, #401303003#] May 2017, thrombolysed. [Angiography, #77343006#] 07/07/2017 - [Severe, #24484000#] disease mLCX and dLCX [Stent, #65818007#] to mLCX to d [LCX, #362036002#]. POBA to dLCX.
IP0277155	CoronaryArteryDisease (Recent [NSTEMI, #401314000#])

VI. RESULT

Table IX shows the results based on the queries we did earlier. We found that there are significantly difference in the results between search by keyword and semantic search even though we did not include the result of negative statement in subsumption, or handle typing errors.

TABLE IX. COMPARISON OF KEYWORD SEARCH AND SEMANTIC SEARCH

Searching Type	Total Record	Negation
Raw	4	1
Codified	33	2

The result from codification using semantic search will be higher than keyword search based on the variation of synonyms and the number of descendants that are returned during subsumption process. While result from semantic search is fewer than keyword search based on negative statements that exist in the records. Typically, the result from semantic search will be higher than keyword search.

VII. CONCLUSION

This paper presents software modules for an ongoing scientific project that supports the automated SNOMED CT mapping of clinical discharge summary data for cardiology queries in clinical facilities. The implemented modules are strictly oriented to data source from clinical facilities in Malaysia.

The scope of the implementation is to improve cardiology query result by observing synonyms, subsumption, and negative statements found in the source data.

Since the source data has been preformatted, data preprocessing is required to remove all formatted tags as it will impact the accuracy of the results.

Subsequently, the codification process is where a matching clinical term is tagged with SNOMED CT. The process begins by breaking those text into each word called as tokenization. Later, those words are combined into phrase for matching and tagging. To improve the accuracy, negation is the process whereby clinical term found in negative statement is excluded.

After the data has been tagged, the source data is ready for query purposes. Semantic search is applied to provide more result than typical query by keyword. In this process, synonym and subsumption result is returned without negative statement. As a result, semantic based queries are significantly better compared to typical query by keyword.

However, typing error is the one major issue that greatly impacts the accuracy, whereby we propose the implementation of auto completion, auto suggestion, or term highlighting to be applied in the Hospital Information System where the source data is input, though it may be costly to implement.

The plan for further development includes testing the accuracy of queries in identifying false positives or negatives, and exploration on how the solution can be applied to different disciplines such as oral health, ophthalmology,etc. Other possibilities such as post coordination and partial matching will be also explored and be included to enrich the results.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All the authors conducted the research, and Abdul Aziz Latip worte the paper, Ma. Stella Tabora Domingo verified the content, Ismat Mohd Sulaiman prepared and analysed the data, respectively.

ACKNOWLEDGMENT

This work was supported by Artificial Intelligence Lab, MIMOS Berhad and Ministry of Health Malaysia. The authors acknowledge, the Artificial Intelligence Lab Director, Dr' Ong Hong Hoe and Deputy Director of Health Informatic Centre Planning Division, Dr Khadzir for their encouragement. Senior Assistant Director, Dr. Syazrin and Dr. Naufal for their assistance on preparing and analysed data.

REFERENCES

[1] S. Boytcheva, "Automatic matching of ICD-10 codes to diagnoses in discharge letters," in *Proceedings of the Workshop on Biomedical Natural Language Processing*, Hissar, 2011, pp. 11-18.

[2] O. Bodenreider, B. Smith, A. Kumar, and A. Burgun, "Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies," *Artificial Intelligence in Medicine*, vol. 39, no.3, pp. 183-195, 2007.

[3] M. J. Lawley, "Exploiting fast classification SNOMED CT for query and integration of health data," in *Proceedings of the 3rd International Conference on Knowledge Representation in Medicine*, Arizona, 2008, pp. 8-14.
[4] S. Boytcheva, "Automatic matching of ICD-10 codes to diagnose in discharges letters," in *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing*, Hissar, 2011, pp. 11-18.
[5] F. Baader and B. Suntrisrivaraporn, "Debugging SNOMED CT using axiom pinpointing in the description logic $\mathcal{EL}^+$," in *Proceedings of the 3rd International Conference on Knowledge Representation in Medicine*, Arizona, 2008, pp. 1-7.

Abdul Aziz Latip has 4 years' experience as Senior Engineer at Artificial Intelligence Lab in MIMOS Berhad. He graduated with Bachelor of Computer Science (Software Engineering) from University Teknologi Malaysia in 2014. He involved on research projects that applied natural language processing for unstructured text and machine learning for image classification.

Ma. Stella Tabora Domingo graduated with BSCoE in the field of computer engineering from University of Baguio, Philippines in 2005. She works currently as a Staff Engineer at the Artificial Intelligence Department, MIMOS Berhad, Kuala Lumpur, Malaysia.

Ismat Mohd Sulaiman has 14 years of experience at the Ministry of Health Malaysia. She graduated as a medical doctor from National University of Malaysia in 2006 and obtained a master's degree in Health Informatics from Karolinska Institute in 2016. She led the MyHarmony project and involved in the planning and implementation of the Malaysian Health Data Warehouse (MyHDW) project. Her team has developed the Malaysian Cardiology SNOMED CT Reference Set that became the standard procedure in developing future SNOMED CT reference set in Malaysia. She also represented Malaysia at the SNOMED CT International from 2011-2019. Other than handling projects, her daily tasks included managing the Health Informatics Standards Unit in the nationwide development and implementation of International Classification of Diseases (ICD), Malaysian Health Data Dictionary (MyHDD), MyHDW Health Reference Data Model (MyHRDM), LOINC, and SNOMED CT. She is now pursuing her Doctoral degree in University of Malaysia in Health informatics. Her current research involves detection and disambiguation of abbreviations in clinical text using artificial intelligence approach. Her other interests are in cardiology, quality registries and quality improvement initiatives.

Tengku Nurulhuda Tengku Abd Rahim received the B.I.T. degree in computer system technology from University of Malaysia Sarawak (UNIMAS), Kota Samarahan, Sarawak, Malaysia in 2000 and the M.C.S. degree in computer science from MARA University of Technology (UiTM), Shah Alam, Selangor, Malaysia in 2016. She is currently a Senior Engineer at Artificial Intelligence Department.