Utilising Bioportal to map terms in the Canada Vigilance Adverse Reaction Online Database to various terminologies and to identify the ideal terminology to supplement MedDRA to aid in Pharmacovigilance activities.

by
Rahul Thandavan
B00755709
Rahul.thandavan@dal.ca

Performed at
Faculty of Computer Science
Dalhousie University 6050 University Ave, Halifax, NS B3H 1W5

Report of Internship for the period May 29 – August 25, 2017
Date Submitted: August 18, 2017

In partial fulfilment of the requirements of the Master of Health Informatics Program, Dalhousie University

Acknowledgement and endorsement

This report has been written by me and has not received any previous academic credit at this or any other institution.

I would like to thank Dr Samuel Stewart and Dr Raza Abidi for their guidance and support throughout the internship.

I would also like to thank my team-mate Soumya Shastri for her contribution throughout the project.

T.Rahul

Rahul Thandavan

Executive summary

The Canada Vigilance Program is Health Canada's post marketing surveillance program that collects and analyzes data related to the adverse events of health products that are marketed in Canada. Health Canada periodically reviews these events and releases safety profile of drugs based on the reported events. The information collected by the Canada Vigilance Program is made available publicly through an online database. This project consists of two main objectives –

1) To map the adverse reaction terms and drug names in the online database to terminologies like SNOMED-CT, MeSH, UMLS and RxNORM and to identify the best approach to map the terms.

2) To identify the best terminology which can help in grouping previously unrelated fine grained MedDRA terms for use in the analysis of adverse events.

The mappings were performed through the Bioportal web service of the National Centre for Biomedical Ontology(NCBO) and then were analysed to determine the accuracy and the coverage of the mappings. The MedDRA terminology that is used for adverse event reporting is a fine-grained terminology so this project focussed on identifying an ideal terminology to group together similar MedDRA terms that are not previously related through MedDRA hierarchy.

This project is highly relevant to Health Informatics as it deals in entirety with healthcare terminologies like SNOMED-CT, UMLS, MeSH, MedDRA etc. which are the basis upon which health information is stored, shared, accessed and analysed. This project was an excellent opportunity to learn a new coding language like python and to learn on using REST-API for mapping terms .It also provided exposure in using large databases. At the end of this internship, the clinical terms were mapped to the identified terminologies and an ideal process to map the terms to various terminologies has been documented. Also, groupings of related MedDRA terms that were not related previously through the MedDRA hierarchy have been identified utilising SNOMED-CT and documented.

The background of the objectives, methods followed and their results, as well as problems and recommended solutions are included in this report.

Table of Contents

1. Introduction

1.1 Utilising Bioportal to map terms in the Canada Vigilance Adverse Reaction Online Database to various terminologies

Bioportal is a Web-based application to access one of the largest repositories of biomedical ontology. It was developed by the National Centre for Biomedical Ontology [1].It constitutes information about most of the terminologies used in healthcare like SNOMED-CT, RxNORM, MeSH, UMLS etc. stored as ontologies. The NCBO has an Annotator web services which use the entity recognizer Mgrep [2], which is the service of interest for this project. The annotator web service processes text and identifies details of matching concepts which exists in the ontologies stored in the Bioportal [2]. The NCBO annotator is available as a Web-based Browser as well as a REST-API that can be programmed to map a given set of terms to a terminology or a group of terminologies. [2]

The Canada Vigilance Adverse Reaction Online Database is maintained as part of the Canada Vigilance Program of Health Canada to keep track of the adverse events relating to health products that are marketed in Canada. [3] This project aims to utilise the Annotator web service of the NCBO through the REST-API to map adverse reaction terms and the drug names in the Canada Vigilance Adverse Reaction Online database to different health care terminologies like SNOMED-CT, RxNORM, MeSH and UMLS.

1.2 Identifying the ideal terminology for grouping related MedDRA terms to aid in Pharmacovigilance

The Canada Vigilance Adverse Reaction Online database is used primarily for identifying causal relationships between a drug and an adverse event. [3] This procedure of identifying the causal relationship between the drug and an event is known as Signal detection. [4] The adverse reaction database is subjected to statistical analysis to identify the relationship between the drug and adverse events. MedDRA being the industrial standard for Adverse event reporting is followed throughout the world for reporting adverse events. [3] It has five levels of hierarchy- 26 System Organ Classes, 332 High-Level Group Terms, 1,688 High-Level Terms, 18,209 Preferred Terms, and 66,587 Low-Level Terms. [3]They are grouped as 26 System organ classes based on the body systems (e.g. Gastrointestinal disorders) and they are comprehensive enough to include most of the adverse events that are reported currently.

Historically MedDRA terms are used alone or their HLT or HLGT is used in identifying signals. But due to the fine-grained nature of MedDRA terms (over 80,000 terms), there is a necessity to group MedDRA terms (like Chest pain and Angina pectoris) which represent similar concepts. [6] Grouping similar MedDRA terms will aid in the better utilisation of the humongous information available in the adverse reaction database. To overcome this shortcoming of MedDRA, this project aims to use the semantically rich information that is available in other

terminologies to group similar MedDRA terms. The mapping of the MedDRA terms to other terminologies completed in the first part of the project will be utilised for this task and the ideal terminology to group MedDRA terms will be explored.

2. Description of the organization

The Canada Vigilance Program is Health Canada's post marketing surveillance program that collects and analyzes data related to the adverse events of health products that are marketed in Canada. Health Canada periodically reviews these events and releases safety profile of drugs based on the reported events. [3] The Canada vigilance program has seven regional offices in the country through which the reports are collected before being forwarded to the National Office for further analysis.[3] The information collected by the Canada Vigilance Program is made available publicly through the online database. The adverse events reported in this database are coded as MedDRA (Medical Dictionary for Regulatory Activities) terms while other information like Drug names and active ingredients are stored as plain text in the database. [3]

This database contains details of adverse events collected by Health Canada regarding adverse events of prescription and non-prescription medications, natural health products, biologics, radiopharmaceuticals and disinfectants since 1965.It is primarily used by Health Canada to evaluate the post market safety profile of the health products and to find out more information about the risk benefit ratio of the products post marketing. [3] The data for the online database is collected through two main mechanism-1) spontaneous surveillance where reporting is done voluntarily by consumers as well as health care practitioners 2) mandatory surveillance- where the pharmaceutical company needs to mandatorily report adverse events concerning their products to the regulatory body. [3] This database has 522,591 reports of adverse events from 05-06-1973 to 30-09-2016.[3]This project is based on the contents of the Canada Vigilance Adverse Reaction Online Database.

3. Description of the work performed

3.1. Utilising the Bioportal REST-API to convert terms in the Canada vigilance adverse reaction online database to various terminologies

3.1.1 Background

MedDRA-The terms for adverse reactions in the database are coded as MedDRA terms.[1] It is not only used for reporting adverse events related to drugs but also for medical devices, vaccines and other health products.[1] It is mainly used by Health Canada for coding adverse events in the Canada vigilance adverse reaction online database and for retrieval and data analysis.[1] MedDRA terminology being specifically designed for adverse events is a very fine grained terminology that is elaborate enough to include most of the adverse reactions that occur for health products and is also being constantly updated to include terms that are not currently present.[3]

MedDRA has five levels of hierarchy- System Organ Classes, High-Level Group Terms, High-Level Terms, Preferred Terms, and Low-Level Terms. The Preferred Term is a single medical concept and Low-Level Terms are synonyms and lexical variants of the Preferred Term. [3]

The MedDRA hierarchy for the term 'Abdominal pain generalized' is depicted below
Example-Low level term(LLT)-Abdominal pain generalized

Preferred Term(PT) – Abdominal pain

High-Level Term (HLT) – Gastrointestinal and Abdominal pain

High-Level Group Term(HLGT) – Gastrointestinal signs and symptoms

System organ class(SOC) – Gastrointestinal disorders

SNOMED-CT -Systematized Nomenclature of Medicine- Clinical Terms is a comprehensive healthcare terminology that covers all fields in healthcare and is an expressive terminology that can help in meaning based data retrieval when used in health records. [7] It is one of the leading healthcare terminologies currently in use by clinical applications like Electronic Health Records throughout the world. SNOMED-CT terminology has concepts which are related through hierarchical relationships and defined relationships.[7] The significant advantage of this terminology is that it can define simple concepts and can also define complex concepts by post coordination of the simple concepts. The SNOMED-CT terminology is a multidimensional classification where terms can be grouped on different criteria and is not as restricted terminology as MedDRA.

MeSH (Medical Subject Headings)-This terminology developed by the National Library of Medicine is a composed vocabulary thesaurus whose main purpose is to index articles, books, magazines and database. [8] It is used primarily for indexing articles for the MEDLINE and

PUBMED database. Each article will be associated with a set of MeSH terms that describe them making the querying process more streamlined. [8] Though this project does not deal with indexing articles this is an attempt to find the feasibility of utilising the rich information of MeSH for a different purpose.

UMLS- "The UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems."[9] UMLS Metathesarus is one of the important tools used in the UMLS and it contains the codes and terms of vocabularies like CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT incorporated in them.[9] All the codes and terms in these different terminologies have a unique UMLS Concept Unique Identifier (CUI) associated with them. The UMLS also has details of the semantic type to which each term belongs to.

RxNORM-It is a terminology mainly used in the pharmacy domain developed by the National Library of Medicine. [10] It contains unique codes for most of the generic and branded drugs and can include details of the route of administration, doses and ingredients. [10]RxNORM is one of the leading terminologies used for drugs and hence it is included in the scope of this project to code drug names present in the database.

3.1.2. Methods

RETRIEVING THE TERMS FROM THE DATABASE -Due to the enormity of the database, this project focusses only on terms in the database relevant to Crohn's disease which accounts for 30890 reports in the database. The terms of adverse reactions and the drugs causing them were extracted from the Canadian ADR database. A total of 3788 terms for adverse reactions and 166 drugs were extracted from the database where the indication for the drug intake was Crohn's disease.

CONVERTING THE TERMS TO DIFFERENT TERMINOLOGIES- The adverse reaction terms found in the database were already MedDRA terms at the level of Preferred Term (PT) in MedDRA hierarchy, but they did not have the relevant MedDRA codes. MedDRA codes for the terms were identified through the Bioportal REST API using Python scripts. The MedDRA code of the immediate parents of the term was also identified and stored as new tables in the database. MedDRA hierarchy information was also included in the following, utilising the hierarchy information of MedDRA available in the UMLS Metathesarus.

RxNORM - RxNORM codes were retrieved from the Bioportal through REST-API for the drugs associated with Crohn's disease.

SNOMED-CT SET-1: Bioportal REST-API was used in utilising the mappings from MedDRA to SNOMED-CT that already existed in Bioportal to map the adverse reaction terms to their relevant SNOMED-CT concepts. The mappings from MedDRA to MeSH and MeSH to SNOMED-CT were also utilised to map the adverse reaction terms to their respective SNOMED-CT terms.

SNOMED-CT SET-2: Another table was created in the database to store separately the information retrieved from Bioportal REST- API where the terms were mapped directly to their SNOMED-CT codes rather than utilising the mappings.

Like the adverse reaction terms, the drug mappings between RxNORM and SNOMED-CT were utilised to identify SNOMED-CT codes . This information along with the SNOMED-CT codes of their parent concepts and the concepts in their hierarchy was stored separately for drugs and reactions. The hierarchy information was included by utilising the hierarchy information of SNOMED-CT in the UMLS Metathesarus.

UMLS-Details of UMLS Concept Unique Identifier was also extracted utilising the Bioportal REST API. Since the UMLS concepts don't have a hierarchical structure, only details of their semantic type were included in the database. But the UMLS had hierarchical information of SNOMED-CT concepts based on their CUI, this hierarchical detail of SNOMED-CT was added to the UMLS CUI that were mapped.

MeSH-The adverse reaction terms from the database were mapped to their respective MeSH codes by utilising the mappings between MedDRA and MeSH while the MeSH code for drugs was extracted from the mapping details between RxNORM and MeSH. Only the mappings between terminologies were utilised to map terms to MeSH and no direct mapping was done for terms in MeSH terminology.

3.1.3. Result

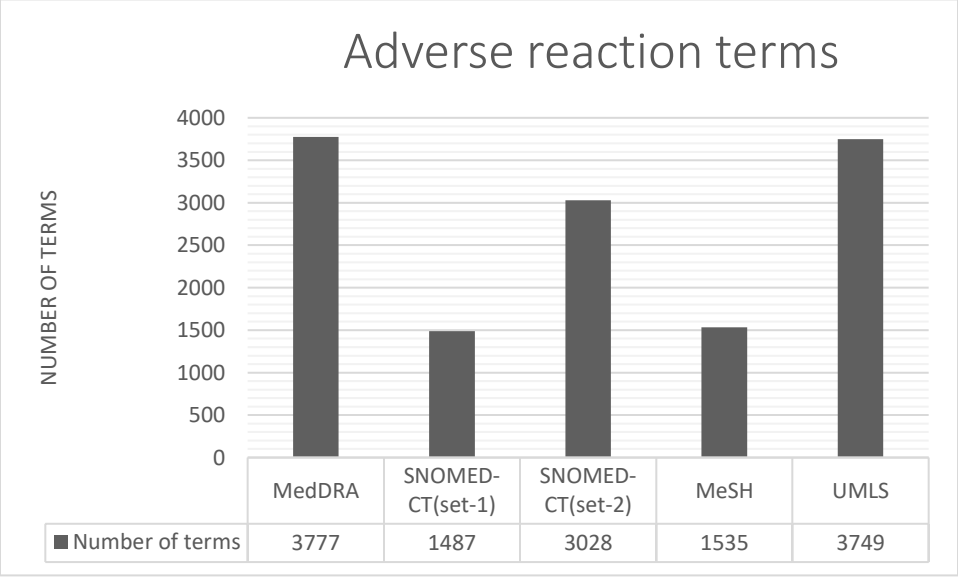


CHART-1 Number of terms that were mapped -terminology wise.

3788 adverse reaction terms were mapped to different terminologies, Chart 1 represents the number of terms that were mapped to different terminologies. For example, the MedDRA term ‘Thrombophlebitis’ (10043570) was mapped to the SNOMED-CT concept ‘Thrombophlebitis’

(64156001), MeSH term ‘Thrombophlebitis’(D013924) and to the UMLS CUI (C0040046) which refers to the concept ‘Thrombophlebitis’.

The high number of terms found in MedDRA 3777 out of 3788(99.7%)can be attributed to the fact that the terms were already standardized MedDRA terms. Since the UMLS CUI was extracted by utilising the CUI found in the JSON file of MedDRA codes[Appendix1], they have an almost similar number of terms 3749 out of 3788(98.9%) like MedDRA. Direct SNOMED-CT mapping of terms (SNOMED-CT SET-2)had the next highest number of terms 3028 out of 3788(79%), while 1487 out of 3788 (39%) SNOMED-CT codes were identified in SNOMED-CT SET-1. MeSH terminology has the second least amount of terms around 40%, this low number of terms may be attributed to the fact that only mappings between MedDRA and MeSH were used to identify the MeSH terms.

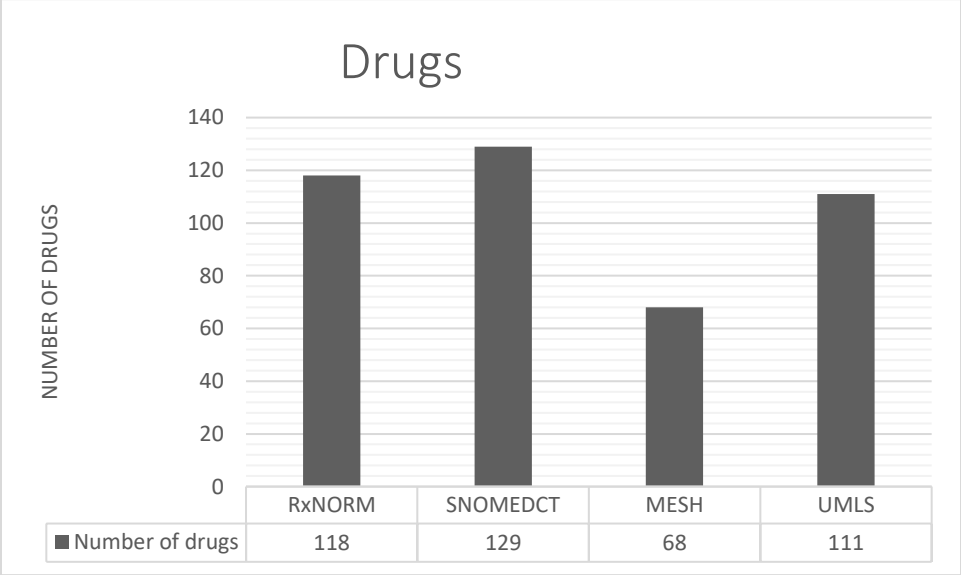


CHART-2- Number of drugs that were mapped terminology wise

166 drugs in total were mapped to concepts in four terminologies RxNORM, SNOMED-CT, UMLS and MeSH. For example, the Drug ‘Vedolizumab’ is mapped to its RxNORM term ‘Vedolizumab’ (1538097), SNOMED-CT term ‘Vedolizumab’ (704257002), MeSH term ‘Vedolizumab’(C543529) and UMLS term ‘Vedolizumab’ with CUI ‘C2742797’.

Chart 2 represents the number of terms that were mapped to the four different terminologies. SNOMED-CT has the highest number of drugs 129 out of 160(77%) that were mapped to their concepts followed by RxNORM 118 out of 160(71%).UMLS has the third highest number of terms 111 out of 160(66%) that were mapped followed by MeSH with 68 out of 160 (41%) which had the least number of terms that were mapped.

A candidate set of 200 adverse reaction terms that were identified randomly from the database, were evaluated manually to determine the accuracy of the mappings of the terms to different terminology.

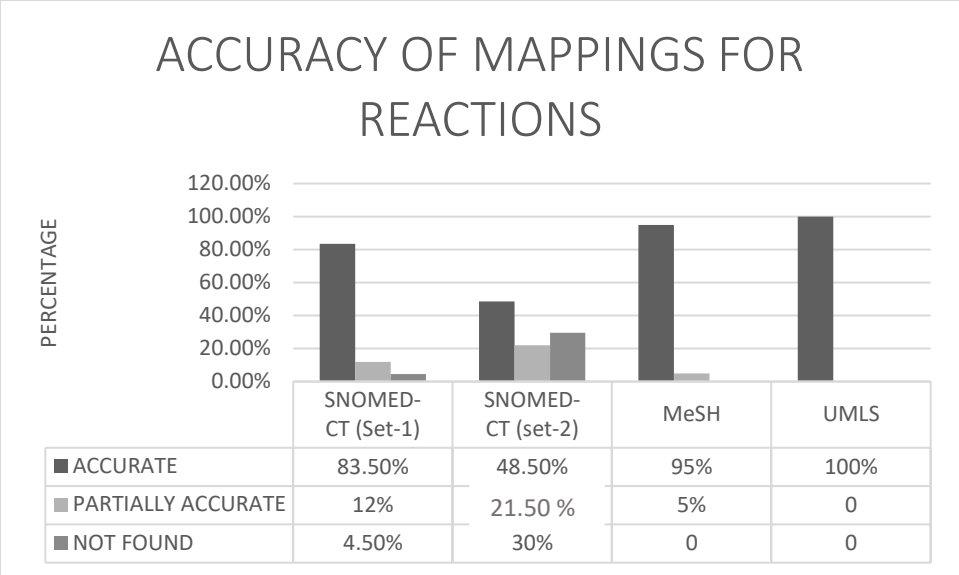


CHART-3 -Accuracy of the mappings for reactions terminology wise

Chart 3 shows the accuracy of the mapped terms based on their terminology. In the sample terms, MedDRA had the highest accuracy of 100%, because the terms were already in MedDRA. Two approaches were taken to map terms in SNOMED-CT first by utilising the mappings from MedDRA to SNOMED-CT (SNOMED-CT SET-1) and the next by directly searching for the terms in SNOMED-CT terminology in the Bioportal (SNOMED-CT SET-2).

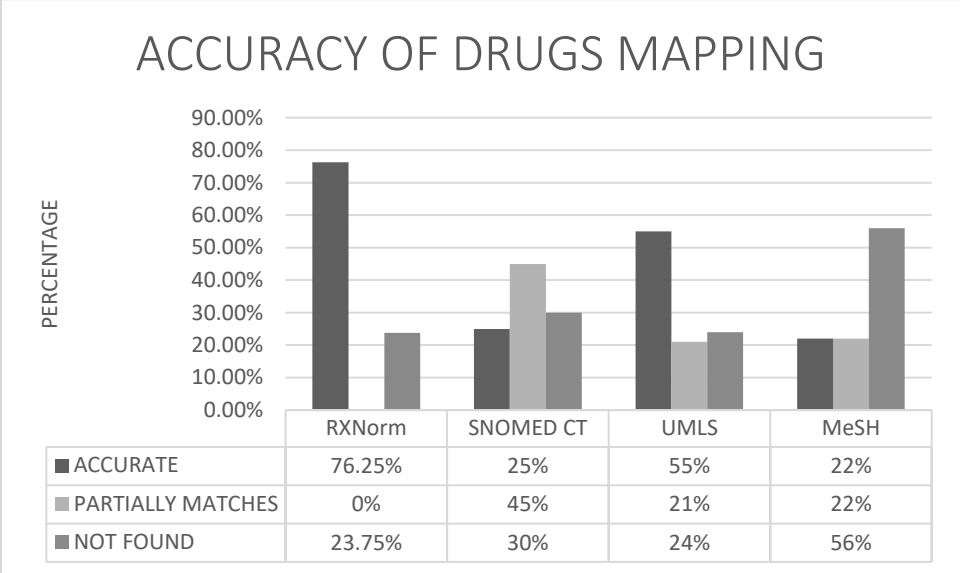
The mappings in SNOMED-CT SET-1 had 83.5% of accurate terms,12% partially accurate and 4.5% incorrect terms. For example, the term '*Lymphadenopathy*' was mapped accurately to the SNOMED-CT term '*Lymphadenopathy*', while the term '*Pneumonia*' was mapped partially accurately to the SNOMED-CT term '*Lobar Pneumonia*'. The term *'Lobar pneumonia'* is a child concept of the concept pneumonia but still the concept *'Pneumonia'* will appear in the hierarchical information of this concept. These kinds of terms which are either children or parent of the original term are described as partially accurate. There are terms that are incorrectly mapped (4.5%) like the term *'Brain neoplasm benign'* which is mapped to the SNOMED-CT concept *'Malignant neoplasm of brain'* in which a benign tumour is mapped to a malignant tumour.

The mappings in SNOMED-CT SET-2 had 48.5% accurate terms,21.5% partial terms and 29.5% terms that were not mapped to their SNOMED-CT concepts. For example, the term *'Nephrostomy'* was mapped accurately to the SNOMED-CT term '*Nephrostomy*' (39834009), while the term *'Meningitis Listeria'* is mapped partially to two SNOMED-CT concepts *'Meningitis'* and *'Listeria'*. Some terms like *'Gastrointestinal disorder'* were not mapped to their SNOMED-CT concept.

The terms that were mapped to MeSH were 95% accurate while the rest 5% terms were partially accurate in which they were mapped to the child or parent concept of the term. Of the 200 terms analysed 100% of them have been mapped correctly to their UMLS CUI. Since the accuracy

analysis is done only for a sample of 200 terms out of 3788 terms further review is necessary to establish the scalability of the accuracy levels established.

Of the 166 Drugs in total associated with Crohn's disease only 160 drugs were taken to evaluate the accuracy of the mappings. The rest 6 were eliminated since many drug names were stored as a single drug in 6 instances.



	RXNorm	SNOMED CT	UMLS	MeSH
ACCURATE	76.25%	25%	55%	22%
PARTIALLY MATCHES	0%	45%	21%	22%
NOT FOUND	23.75%	30%	24%	56%

CHART-4 -Accuracy of the mappings for drugs

Among the 160 drug names, 122 out of 160 (76.25%) of the drugs were exactly mapped to their relevant RxNORM concepts. For the SNOMED-CT 40 out of 160 (25 %) of the terms that were mapped were exact matches, 72 (45%) drugs were partial matches and 48(30%) drugs were not found.Drugs with partial matches refer to a situation where a drug '*Apo methotrexate'* was mapped to the SNOMED-CT concept '*Methotrexate*' alone. The UMLS had slightly better accuracy than RxNORM and SNOMED-CT with 87 drugs (55%) accurately mapped,34 drugs (21%) partial mapped and 39 (24%) drugs not found. MeSH was the least accurate with 35(22%)drugs accurate mapped, 35(22%) drugs partially mapped and 90(56%) drugs were not found.

3.2 Identifying the ideal terminology for grouping related MedDRA terms to aid in Pharmacovigilance

3.2.1. Background

The beginnings of Pharmacovigilance can be traced back to the Thalidomide disaster of 1961 after which the World Health Organization realized the need for a system to track down previously unknown or poorly understood adverse events of health products [11].WHO defines pharmacovigilance as **"the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem."**[11]To serve this purpose of monitoring the safety of drugs Health Canada collects all data related to adverse events in Canada and stores them in the Canada vigilance adverse reaction online database for further analysis.

The main goal of this task is to identify the terminology that can produce number of grouping of previously unrelated MedDRA terms to facilitate accurate data retrieval for signal detection. A MedDRA term was identified and then the MedDRA hierarchy of the 3777 terms was searched to identify other terms that had the same concept at any level of the hierarchy.

For example, for the MedDRA term '*Abdominal Pain'* a search was conducted to identify other MedDRA terms which had the word Abdominal Pain in their hierarchy. Six new MedDRA terms related to *'Abdominal pain'* were identified- *'Abdominal Pain Upper', 'Abdominal Pain Lower', 'Gastrointestinal Pain', 'Abdominal rigidity', 'Oesophageal pain'* and *'Abdominal tenderness'*.

These 6 terms have the word Abdominal pain in their hierarchy at the level of MedDRA High-Level term *('Gastrointestinal and abdominal pains')* and this grouping of terms is readily available in MedDRA for the regulatory authorities, to be used for statistical analysis to identify the causal relationship between a drug and the reaction 'Abdominal pain'. MedDRA terminology is very specialised and it is rich enough to define most of the possible adverse reaction terms, but due to its strict hierarchy, not all the related terms are grouped together.

There are also other terms which are synonyms of Abdominal pain like '*Flank pai*n' (Flank is the region of the abdomen on the sides) which are grouped not only under different HLT, but also different SOC. In the case of Flank pain, the hierarchy is shown in the chart 6. It comes under the SOC general disorders and administrative site conditions, due to the multiaxiality of MedDRA the same Flank pain can also be seen under two other SOC also *'Renal and urinary disorders'* and *'Musculoskeletal and connective tissue disorders'* but no where in any hierarchy it has *'Gastrointestinal and abdominal pains'* High-Level term. So, the number of reactions with the term 'flank pain' would have been missed when searching for adverse reaction relating to *'Abdominal pain'*. If we go into the hierarchical details of the concept Flank pain in SNOMED-CT as shown in chart 7, the term '*Flank pain'* is stored as child concept of *'Abdominal Pain'*.

Since SNOMED-CT doesn't have a strict hierarchy it can have more relationships between concepts than MedDRA.

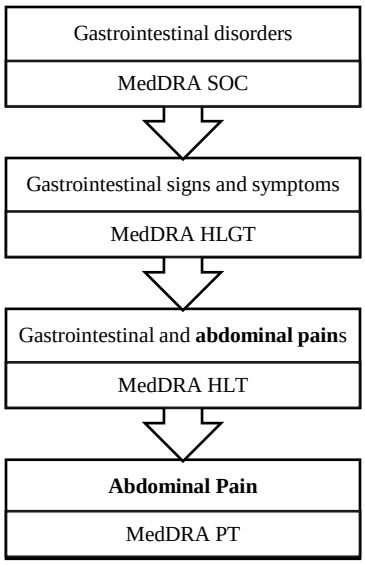


Chart 5 MedDRA hierarchy of 'Abdominal Pain'

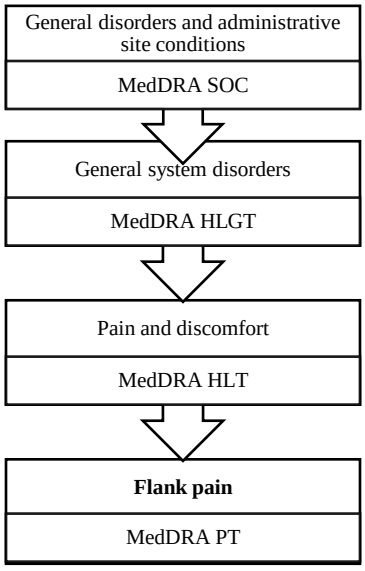


Chart 6 MedDRA hierarchy of 'Flank pain'

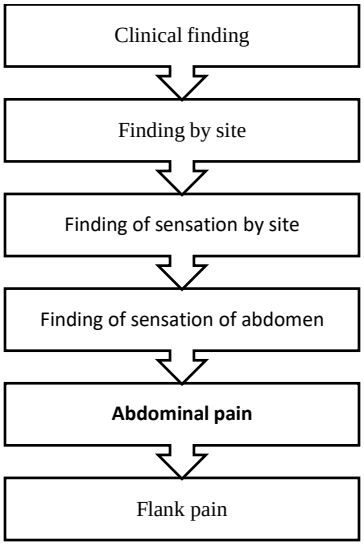


Chart 7 SNOMED-CT hierarchy of 'Flank pain'

MedDRA has a grouping of terms stored as Standardized MedDRA queries(SMQ) for commonly occurring adverse events to be used by the regulatory bodies for signal detection in Pharmacovigilance. But this SMQ does not cover all the adverse events that are being studied by

the regulatory bodies and they are constantly being developed manually after extensive consultation with all the stakeholders.

The main objective of this project is to find a similar grouping of MedDRA terms utilising other terminologies like SNOMED-CT, UMLS and MeSH and to identify the ideal terminology that can create a maximum number of groupings of previously unrelated MedDRA terms to be used in Pharmacovigilance and to aid in formulating Standardized MedDRA queries.

3.2.2. Methods

The adverse reaction MedDRA terms which were present in the database relevant to Crohn's disease were mapped to the relevant SNOMED-CT, MeSH and UMLS concepts along with the hierarchical information in the first part of the project. To group terms with a similar concept, a MedDRA term was selected and the hierarchal information of all other terms in the database based on other terminologies was searched to identify MedDRA terms that were previously unrelated.

Groupings were analysed separately for both sets of SNOMED-CT concepts, SNOMED-CT SET-1 was identified based on the mappings between MedDRA and SNOMED-CT while the SNOMED-CT SET-2 was identified based on direct mapping to SNOMED-CT concepts using Bioportal. Since UMLS terms don't have a hierarchy of their own, SNOMED-CT hierarchy information for UMLS CUI's that were available in the UMLS Metathesarus was used. This data was utilised to identify the groupings of MedDRA terms that can be formulated utilising other terminologies and they were compared with groupings created utilising MedDRA to identify concepts that were left out in MedDRA.

SAMPLE-800 terms (~ 20% of the total 3788) were taken as sample terms. The adverse reaction terms were ordered based on the number of instances they occur in the database and the sample terms included top 400 terms with the highest count,200 terms with a count of 5-7 instances and 200 terms with a count of 1 instance in the database.

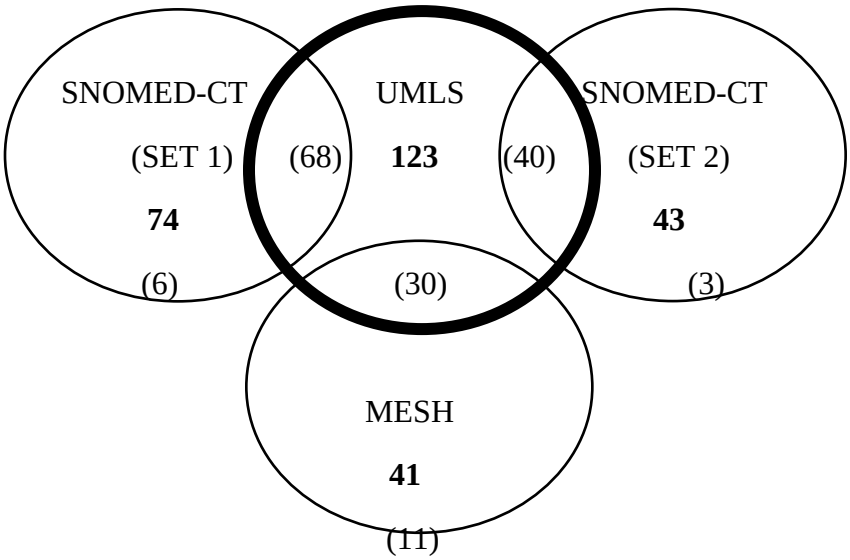
3.2.3. Results

Table-1 provides details on the number of grouping of MedDRA terms which were previously unrelated but now identified using other terminologies. A total number of 128 groupings of adverse reaction terms were found on using the UMLS concepts along with SNOMED-CT hierarchy [Appendix-3], which was followed by the SNOMED-CT SET-1 (74) which utilised mappings between MedDRA and SNOMED-CT. The SNOMED-CT SET-2 could identify 43 grouping of terms while MeSH had the least number of groupings at 41.

TERMINOLOGY	Extra grouping of terms found out of 800 terms
SNOMED-CT *(*SNOMED-CT SET-1*)*	74
SNOMED-CT *(*SNOMED-CT SET-2*)*	43
MESH	41
UMLS	123

Table-1-Number of new groupings of terms found terminology wise

CHART-8-Overlap of the grouping of MedDRA terms found based on different terminologies with the UMLS approach.



For example, for the term 'Abdominal Pain' the search in MedDRA revealed six new MedDRA terms related to *'Abdominal pain' -'Abdominal Pain Upper', 'Abdominal Pain Lower', 'Gastrointestinal Pain', 'Abdominal rigidity', 'Oesophageal pain'* and *'Abdominal tenderness'*. While on searching in the SNOMED-CT hierarchy (SNOMED-CT SET-1&2) two new MedDRA terms with similar meaning were identified-'*Renal Colic*' and '*Flank pain'*. On utilising the details from the MeSH terminology no new terms were identified, while on utilising the SNOMED-CT hierarchy details based on the UMLS CUI, 11 new relevant terms were identified- *'Biliary colic', 'Bladder pain', 'Flank pain', 'Gallbladder pain', 'Groin pain', 'Hepatic pain' , 'Proctalgia', 'Renal colic', 'Renal pain', 'Suprapubic pain'* and *'Uterine pain'*. The new terms identified through SNOMED-CT SET 1&2 were also present in the grouping identified through the UMLS.

MedDRA terms	MedDRA- System Organ Class (SOC)
'Proctalgia'	'Gastrointestinal disorders'
'Groin pain'	'Musculoskeletal and connective tissue disorders'
'Flank pain'	'Musculoskeletal and connective tissue disorders'
'Gallbladder pain'	'Hepatobiliary disorders'
'Hepatic pain'	'Hepatobiliary disorders'
'Biliary colic'	'Hepatobiliary disorders'
'Renal colic'	'Renal and urinary disorders'
'Renal pain'	'Renal and urinary disorders'
'Bladder pain'	'Renal and urinary disorders'
'Suprapubic pain'	'General disorders and administration site conditions'
'Uterine pain'	'Reproductive system and breast disorders'

Table 2-New MedDRA terms identified along with their System Organ Class(SOC)

Among the new terms identified only one term belongs to the MedDRA SOC 'Gastrointestinal disorders' while others belong to different SOC's as shown in Table-2. By grouping MedDRA terms based on SNOMED-CT hierarchy, it is evident that it is possible to group similar terms in MedDRA that occur in the different hierarchy that were previously unrelated. When all these 11 newly identified MedDRA terms were included in the query to calculate the number of instances the reaction 'Abdominal Pain' has occurred in the database restricted to diagnosis of 'Crohn's disease', the count increased by **379** from **5651** (while using the term grouping formed through MedDRA) to **6030** (while using the MedDRA term grouping along with the group of terms formed by SNOMED-CT)**.**

Of the different terminologies used the UMLS CUI along with SNOMED-CT hierarchy produced the highest number of groupings of terms and the quality of the groupings was far superior in the UMLS than the other approaches. The chart-8 shows the overlap of the groupings that were formed by SNOMED-CT (SET-1&2), MeSH and UMLS. It is clearly evident from the Chart-8 that the groupings formed with UMLS CUI cover 91% of the groupings in SNOMED-T set 1, 93% of the groupings found in SNOMED-CT set 2 and 73% of the grouping found in grouping of terms in MeSH.

Based on the results of the number of groupings, SNOMED-CT will clearly be the ideal candidate to be used for supplementing MedDRA in grouping of MedDRA terms in relevance to this database.

4. Relation to Health Informatics

This Project is very much related to Health Informatics as it deals with the Canadian Adverse Reaction Online database which itself is a perfect example for the meeting point of healthcare and technology where clinical reports are stored in a computer database. This project also provided exposure to different medical terminology standards like SNOMED-CT, RxNORM, UMLS, MeSH and MedDRA which are highly relevant to the health informatics discipline.

Most of the project is based on the MySQL database in which the Canadian ADR database was accessed and manipulated. The Health Informatics course has provided the author with the skills that were required to use the database. Throughout the length of the project, the author was able to improve his knowledge about databases by manipulating the data in the adverse reaction database , importing data from different sources and adding new tables to the existing database. This project was an excellent opportunity to learn about accessing REST-API through Python, as a significant portion of the project is based on the data collected from the Bioportal REST-API through Python. The skills in PHP learned during the course were very helpful in learning a new coding language like Python.

This project also provided the author with the opportunity to map clinical concepts to different healthcare terminologies utilising Bioportal REST-API which is highly relevant to Health Informatics as most of the problems that occur in today's health informatics field are related to interoperability issues related to terminologies.

5. Problems and solutions

During this project, one of the significant problems that came up was during the mapping of the adverse reaction terms that were in MedDRA to the SNOMED-CT terminology. MedDRA and SNOMED-CT are terminologies with their own specialization the former has a standard 5 level hierarchy structure while SNOMED-CT does not have a standard hierarchal structure and may vary from term to term.

During the mapping process, 3777 terms were identified in MedDRA, since the Bioportal had details of mappings between MedDRA and SNOMED-CT it was utilised to map the terms in MedDRA to their respective SNOMED-CT codes. But only 1487 terms(39%) were mapped to their SNOMED-CT codes while the rest of the 2290 MedDRA terms did not have pre-existing mappings to SNOMED-CT in the Bioportal.

To overcome this defect, the author tried to map the adverse reaction terms in the database directly to their SNOMED-CT codes through the Bioportal Annotator. Through this approach, SNOMED-CT concepts were found for 3028 adverse reaction terms out of the total of 3788. But

after the accuracy of the mappings were evaluated for a sample of 200 terms it was clear that this approach mapped less than 50% (48.5%) terms accurately while some were not found and for others, the adverse reaction term was split and each word was assigned a separate SNOMED-CT code. For example, for the MedDRA term *'Blood Immunoglobulin M increased'* Bioportal returned three different SNOMED-CT codes -

SNOMED-CT CODES	TERM
87612001	'Blood'
74889000	'Immunoglobulin M'
35105006	'Increased'

This might be attributed to the way SNOMED-CT is designed to form complex concepts by combining multiple simple concepts through compositional grammar. The Bioportal annotator cannot form SNOMED-CT concepts as compositional grammar since it is based on text recognition where every term is searched in the respective ontology and if there are no full matching terms the respective code of the all the terms are delivered.

The next approach was to map the adverse reaction terms to SNOMED-CT by using the UMLS Concept Unique Identifier (CUI) of the MedDRA terms and to find the SNOMED-CT code for the identified CUI in the UMLS Metathesarus. In the sample terms that were evaluated for accuracy, UMLS CUI was the most accurate with 100 % of the CUI identified being accurate. Through this approach, 66%(2563 out of 3788) terms were mapped to their SNOMED-CT codes and their hierarchy. This result is in line with the research conducted by Bodenreider et al where it was established that around 55.5% of terms in MedDRA had mappings to SNOMED-CT through UMLS. [12] The relatively higher number of mappings in our result can be attributed to the fact that Bodenreider et al conducted analysis on all the terms in MedDRA[12], where in this project only around 3788 terms that were associated with Crohn's disease was analysed and mapped.

Though none of the approaches used in this project could map accurately all the MedDRA terms to SNOMED-CT, the approach using UMLS CUI to map the SNOMED-CT codes is the most accurate approach to map the MedDRA terms in this database. Even though it was only able to identify 66% of the SNOMED-CT codes when compared to the direct SNOMED-CT approach which was able to map 79% terms. The UMLS approach in the sample had 100% accuracy while the SNOMED-CT direct mapping was only able to map 48.5% accurately. Also, the approach in which SNOMED-CT concepts were identified based on the mappings in Bioportal between MedDRA and SNOMED-CT could map only 39% of the total terms while the UMLS approach was able to map 66%. Though the sample used for the evaluation of the accuracy is small, further analysis should be done to prove the efficacy of this approach in relevance to the terms in the database.

Another problem that was encountered during this project is related to the grouping of MedDRA terms that were formed through SNOMED-CT. Most of the terms in the groupings that were formed based on the SNOMED-CT hierarchal structure were relevant but there were also

some instances where unrelated terms were grouped. In one particular example for an adverse event term 'Tic' 338 new MedDRA terms were grouped together based on the SNOMED-CT while 752 MedDRA terms were grouped based on MedDRA hierarchy ,In this case not all the terms that were grouped had an inherent meaning of Tic (sudden repetitive movements involving discrete muscle groups) but most of the words that were grouped had words with the letters 'Tic' like Neoplas***tic*** and trauma***tic*** in them .This is attributed to the fact that the query to group terms identified all the words with those letters.This is one of the main draw backs of this approach and hence manual checking of the newly formed grouping of MedDRA terms is highly warranted before using the groupings for statistical analysis.

6. Conclusions and Recommendations

With the widespread adoption of SNOMED-CT across the clinical domain, mapping the MedDRA terms to the SNOMED-CT codes will facilitate a common data model that will contribute to the seamless data movement between the clinical applications like EHR and adverse reaction databases, improving both the quality and quantity of the data in the adverse reaction database. With only a fraction of cases of adverse events being reported any process that can streamline the data migration from the different clinical application will be very significant in identifying adverse events related to drugs.

Bioportal served as an invaluable tool to map the adverse reactions MedDRA terms and drugs in the Canada Vigilance Adverse Reaction Online Database to various terminologies including SNOMED-CT. In the due course of the project, it was established that mapping the terms to SNOMED-CT terminology will be more beneficial owing to the rich semantic information and granularity available in the SNOMED-CT terminology. With the SNOMED-CT being a subset of the UMLS terminology and a high level of UMLS CUI mapped to the MedDRA terms, mapping MedDRA terms to the UMLS was identified as a comprehensive approach to map the terms with the high level of accuracy. The MeSH terminology being a standard used for bibliographic indexes obviously was not as comprehensive enough as SNOMED-CT in relevance to the terms as well as the functionalities of this database.

The quality of mappings in Bioportal and UMLS Metathesarus between MedDRA and SNOMED-CT had a direct effect on this project. With the mappings between terminologies being constantly updated, in the future it might be able to seamlessly integrate systems that work in MedDRA and SNOMED-CT terminologies. Regarding the grouping of MedDRA terms, it has been established that SNOMED-CT will be the terminology of choice to supplement MedDRA in grouping adverse reaction terms due to its high number as well as the quality of grouping of the MedDRA terms.

Based on the currently available data on mappings the best way to map the MedDRA terms to their respective SNOMED-CT will be to map the terms in MedDRA to their respective UMLS Concept Unique Identifiers using Bioportal web services and then use the UMLS Metathesarus which is a melting point of different healthcare terminologies to identify the relevant SNOMED-CT codes based on the UMLS CUI .Even though in this project this approach has lead to the

mapping of only 66% of MedDRA terms, and there are other approaches that boast of a significantly higher number of mappings, this approach has been identified as the one with the most accurate level of mappings.

To further continue with this project the Python scripts to map the terms in the database through Bioportal and the UMLS Metathesarus database are scalable and hence can be used to map the entire terms in the database in the future. With all the adverse reaction terms in the database mapped to their respective SNOMED-CT codes, the quality of groupings is also expected to improve substantially.

Reference

[1] Canada, H. (2017). *Canadian Adverse Drug Reaction Monitoring Program - Canada.ca*. *Canada.ca*. Retrieved 8 August 2017, from https://www.canada.ca/en/health-canada/corporate/about-health-canada/activities-responsibilities/access-information-privacy/canadian-adverse-drug-reaction-monitoring-program.html

[2] Whetzel, P., Noy, N., Shah, N., Alexander, P., Nyulas, C., Tudorache, T., & Musen, M. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, *39*(suppl), W541-W545. http://dx.doi.org/10.1093/nar/gkr469

[3] Canada, H. (2017). *About the Medical Dictionary for Regulatory Activities - Canada.ca*. *Canada.ca*. Retrieved 8 August 2017, from https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/adverse-reaction-database/about-medical-dictionary-regulatory-activities-canada-vigilance-adverse-reaction-online-database.html

[4] Meyboom, R. H., Egberts, A. C., Edwards, I. R., Hekster, Y. A., de Koning, F. H., & Gribnau, F. W. (1997). Principles of signal detection in pharmacovigilance. *Drug safety*, *16*(6), 355-365.

[5] Whetzel, P. L. (2013). NCBO Technology: Powering semantically aware applications. *Journal of biomedical semantics*, *4*(1), S8.

[6] Dupuch, M., & Grabar, N. (2015). Semantic distance-based creation of clusters of pharmacovigilance terms and their evaluation. *Journal of biomedical informatics*, *54*, 174-185.

[7] *SNOMED International*. (2017). *Snomed.org*. Retrieved 8 August 2017, from http://www.snomed.org/snomed-ct/what-is-snomed-ct/how-does-snomed-ct-work

[8] *Medical Subject Headings (MESH®) Fact Sheet*. (2017). *Nlm.nih.gov*. Retrieved 8 August 2017, from https://www.nlm.nih.gov/pubs/factsheets/mesh.html

[9] *UMLS Quick Start Guide*. (2017). *Nlm.nih.gov*. Retrieved 8 August 2017, from https://www.nlm.nih.gov/research/umls/quickstart.html

[10] RxNorm Overview. (2017). Nlm.nih.gov. Retrieved 8 August 2017, from https://www.nlm.nih.gov/research/umls/rxnorm/overview.html

[11] *Pharmacovigilance*. (2017). *World Health Organization*. Retrieved 8 August 2017, from http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/

[12]Bodenreider, O. (2009). Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. In *AMIA Annual Symposium Proceedings* (Vol. 2009, p. 45). American Medical Informatics Association.

[13]Alecu, I., Bousquet, C., Mougin, F., & Jaulent, M. C. (2005). Mapping of the WHO-ART terminology on Snomed CT to improve grouping of related adverse drug reactions. *Studies in health technology and informatics*, *124*, 833-838.

Appendix

Appendix-1

JSON-file of MedDRA term Abdominal pain retrieved through Bioportal


```
{
   prefLabel: "Abdominal pain",
   synonym: [ ],
   definition: [ ],
 - cui: [
      "C0000737"
   ],
 - semanticType: [
      http://purl.bioontology.org/ontology/STY/T184
   ],
   obsolete: false,
   @id: http://purl.bioontology.org/ontology/MEDDRA/10000081,
   @type: http://www.w3.org/2002/07/owl#Class,
 - links: {
      self: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081,
      ontology: http://data.bioontology.org/ontologies/MEDDRA,
      children: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/children,
      parents: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/parents,
      descendants: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/descendants,
      ancestors: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/ancestors,
      instances: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/instances,
      tree: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/tree,
      notes: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/notes,
      mappings: http://data.bioontology.org/ontologies/MEDDRA/classes/http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081/mappings,
      ui: http://bioportal.bioontology.org/ontologies/MEDDRA?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FMEDDRA%2F10000081,
    - @context: {
         self: http://www.w3.org/2002/07/owl#Class,
         ontology: http://data.bioontology.org/metadata/Ontology,
         children: http://www.w3.org/2002/07/owl#Class,
         parents: http://www.w3.org/2002/07/owl#Class,
         descendants: http://www.w3.org/2002/07/owl#Class,
         ancestors: http://www.w3.org/2002/07/owl#Class,
         instances: http://data.bioontology.org/metadata/Instance,
         tree: http://www.w3.org/2002/07/owl#Class,
         notes: http://data.bioontology.org/metadata/Note,
         mappings: http://data.bioontology.org/metadata/Mapping,
         ui: http://www.w3.org/2002/07/owl#Class
      }
```

Appendix-2

Python script used for mapping terms in Bioportal

```
REST_URL = "http://data.bioontology.org"
API_KEY = "a8710789-e74c-4842-98bc-393d2de46397"

def get_json(url):
   opener = urllib.request.build_opener()
   opener.addheaders = [('Authorization', 'apikey token=' + API_KEY)]
   return json.loads(opener.open(url).read())

def print_annotations(annotations, get_class=True):
   values=""
   for result in annotations:
       class_detail1 =get_json(result["annotatedClass"]["links"]["self"]) if get_class else result["annotatedClass"]
       class1=class_detail1
       if len(class1)<1 :
           print("not found")
           return
       strTemp1=class1["@id"].rsplit('/',1)
       strTemp2=class1["@id"].rsplit('/',2)
       print(json.dumps(["codeID :" +strTemp1[1], "Preferred Name: " + class1["prefLabel"], "Terminology:" + strTemp2[1]],inde
       values= values+ '('+strTemp1[1] +',"'+e+'"'+','+'"'+class1["prefLabel"]+'"),'
   return(values)

with open('ADR_term.json') as json_data:
    d = json.load(json_data)
    query = "INSERT INTO MEDDRAterm(meddraCUID,term) VALUES "
    cnx = pymysql.connect(user='adr',password='adrPASSWORD',host='129.173.20.222',database='canadianADR')
    cursor = cnx.cursor()
    counter = 0
    values = ""
    for strVal in d:
        text_to_annotate = str(strVal).split(",")[0]
        e=str(strVal).split(",")[0]
        #d=e.rsplit('[',1)[1]
        print(e)
        annotations = get_json(REST_URL + "/annotator?text=" + urllib.request.quote(text_to_annotate) + "&ontologies=MEDDRA&lo
        values = values + print_annotations(annotations)
        counter += 1
        if counter == 4:break
    values= values[::-1].replace(",","",1)[::-1]
    print(values)
    cursor.execute(query+values)
    cnx.commit()
    print(query+values)

    cursor.close()
    cnx.close()
```

Appendix-3

Sample of New groupings of terms found through UMLS and SNOMED-CT along with Grouping of terms found in MedDRA and the variations in the number of reactions in the database.

Term	Grouping of terms from MedDRA	Old Reaction Count	Difference	New reaction count	Grouping of terms extra in SNOMED-CT
1)Effusion	'Effusion'	167	**114**	281	'Bloody discharge'
	'Joint effusion'				'Palindromic rheumatism'
	'Malignant pleural effusion'				'Purulent discharge'
	'Middle ear effusion'				'Scab'
	'Pericardial effusion'				
	'Pleural effusion'				
	'Pneumothorax spontaneous'				
	'Pneumothorax traumatic'				
	'Pneumothorax'				
2)Rectal abscess	'Perirectal abscess'	393	**612**	1005	'Anal abscess'
	'Rectal abscess'				
3)Musculoskeletal pain	'Musculoskeletal pain'	161	**2186**	2347	'Arthralgia'
					'Bone pain'
					'Coccydynia'
					'Facet joint syndrome'
					'Musculoskeletal chest pain'
					'Myalgia'
					'Occipital neuralgia'
					'Spinal pain'
					'Tendon pain'
4)Inflammatory bowel disease	'Inflammatory bowel disease'	51	**266**	317	'Colitis ulcerative'
5)Muscle abscess	'Muscle abscess'	6	**17**	23	'Psoas abscess'