A High Throughput Semantic Concept Frequency Based Approach for Patient Identification: A Case Study Using Type 2 Diabetes Mellitus Clinical Notes

Wei-Qi Wei[1, 2, 3] MM, Cui Tao[1]PhD, Guoqian Jiang[1]PhD, and Christopher G. Chute[1] MD, DrPH

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN
[2]Health Informatics Program, University of Minnesota Medical School. Minneapolis, MN
[3]BICB Traineeship Program, University of Minnesota, Rochester, MN

Abstract: *Current research on high throughput identification of patients with a specific phenotype is in its infancy. There is an urgent need to develop a general automatic approach for patient identification. Objective: We took advantage of Mayo Clinic electronic clinical notes and proposed a novel method of combining NLP, machine learning, and ontology for automatic patient identification. We also investigated the benefits of involving existing SNOMED semantic knowledge in a patient identification task. Methods: the SVM algorithm was applied on SNOMED concept units extracted from T2DM case/control clinical notes. Precision, recall, and F-score were calculated to evaluate the performance. Results: This approach achieved an F-score of above 0.950 for both groups when using all identified concept units as features. Concept units from semantic type—Disease or Syndrome contain the most important information for patient identification. Our results also implied that the coarse level concepts contain enough information to classify T2DM cases/controls.*

Introduction and Background:

The combination of DNA biorepositories with phenotype information for large-scale, high throughput genetic research will enable further exploration of how genetic variation contributes to personal health, disease, and treatment[1, 2]. To conduct a successful genome-wide association (GWA) study, a significant number of subjects (cases and controls) are often required. Without adequate subjects, a GWA study may not carry sufficient statistical power and researchers may not be able to derive conclusive results.

Current research on high throughput identification of patients with a specific phenotype is in its infancy. The manual chart review process of patients' medical records is extremely labor and resource intensive, which is hardly affordable for a large volume patient identification task. Although computer systems are routinely used for clinical data storage and analysis, tremendous human efforts are required for gathering, abstracting, and reviewing a large volume of patients' charts. A manual chart review process also takes a long time to complete. Domain-expert proposed algorithms may leverage structured electronic medical data, such as diagnose codes and lab test results, which can obviate human abstraction and improve the efficiency of identification. Previous studies have demonstrated that using diagnosis codes alone is not able to provide quality case-finding results[3-5]. It is necessary to find other reliable resources for a general patient selection. Both manual chart review and domain-expert proposed algorithms are created for a specific use case. In addition, they must be optimized by patient-care domain experts for terminology coding and clinical documentation[6]. Domain experts play major roles in both methods. Thus there remains the possibility of error in either modality.

There is an urgent call for developing a general automatic approach for patient case-finding and characterization. In this study, we took advantage of historical Mayo Clinic electronic clinical notes and proposed a novel general automatic approach for disease phenotyping patients. The approach combines Natural Language Processing (NLP), machine learning based classification algorithms, and semantic techniques. We were also interested in discovering if any existing semantic knowledge is helpful for preselecting the most important features to a specific phenotype, and mitigating the computational burden due to significant amount of clinical data, thereby improving the performance of the case identification.

Clinical notes, ranging in length from a few sentences to several pages, contain a significant amount of complex and detailed clinical phenotype information. This information constitutes the primary target for case identification[6]. Clinical notes are rich data sources for general patient identification. A main obstacle restricting the usage of clinical notes is their unstructured format, which makes it complicated to search, aggregate, and analyze. NLP provides a possibility to bridge the gap between clinical free text and structured data, allowing humans to deal with a familiar natural language while enabling machines to effectively process data[7, 8]. Although current NLP techniques may not fully express the knowledge and relationships within a context, they are successful at automated recognition of biomedical named entities (NEs): diseases/disorders, signs/symptoms, anatomical sites, drugs, and procedures. A Mayo Clinic NLP package, called clinical Text Analysis and Knowledge Extraction System (cTAKES)[9], offers the public an application programming interface (API) to identify NEs from clinical notes. The identified NEs can be mapped to unique concept identifiers in an appropriate terminology. This package has been successfully used in a population-based cohort study of congestive heart failure[10, 11].

Machine learning based classification algorithms are good at automatically learning to recognize complex patterns and making intelligent decisions based on data. Many algorithms have been widely used and proven successful in many practical problems[12]. We used the support vector machine (SVM) algorithm in this study because SVM is efficient with large amounts of data[13]. Feature selection, which is commonly used in machine learning, is the technique of selecting a subset of candidate features for building robust learning models. It can significantly improve the performance of learning models by alleviating the curse of dimensionality, enhancing generalization capability, speeding up learning process, and improving model interpretability[13, 14]. However, when a large number of features are available, it is impractical to find an optimal subset of features because it requires an exhaustive search of all possible subsets of the chosen cardinality.

Ontology enable the appropriate and advantageous formalization of knowledge, and thus may potentially be helpful for finding the most related information of a target phenotype, and improving the performance of patient identification. Specifically, we chose SNOMED CT in this study because it is considered the most comprehensive clinically oriented healthcare terminology available. One of the major benefits of using SNOMED CT is its high content coverage[15-17]. In the past 40 years, SNOMED CT has been widely used for clinically related applications and research[18].

The feasibility of our proposed approach was evaluated by using clinical notes of Type 2 Diabetes Mellitus (T2DM) cases/controls.

Methods:

Data Collection: We chose Mayo Clinic's medical data from the 2007 calendar year and focused upon patients' clinical notes. We also restricted this study to Mayo patients from Minnesota's Olmsted County to provide a population-based context. This increases the likelihood that the subjects were receiving primary care at Mayo Clinic. All patients without Minnesota Research Authorization consent were excluded. With protocols approved by the Mayo Clinic IRB, the qualified patients were screened first by applying the Northwestern University (NW) T2DM algorithm[19]. The NW algorithm uses diagnostic codes, lab tests, and medication data to build up the inclusion and the exclusion criteria. It has been validated across institutes including the Mayo Clinic and Vanderbilt University Medical Center. We also checked the cases based on previous experience and expertise with manual review of patients' detailed medical records.

A total of 1,600 T2DM patients were identified from the clinical data from 2007. A total of 1,600 controls, matched with cases by age and gender, were randomly chosen from the remaining population. The average age of selected subjects was 63.5±13.5 (mean±STD). The gender ratio was 1.2 male(s)/female. For the T2DM case group, a patient's average number of clinical notes in 2007 was 21.0±22.6 (mean±STD); for the control group, the number was 15.1±20.0. Statistical analysis indicated there was a significant difference between the numbers of notes available between the two groups (P<0.01). This may because T2DM increases patients' risk for many serious health problems. Many patients with T2DM are with other complications, such as eyes, foot, and skin. Thus, their average number of clinical notes is higher than others.

All of the subjects' clinical notes were processed using cTAKEs to extract distinct SNOMED CT concept units. A concept unit is defined as a concept with a negation value (positive or negative). For example, concept "diabetes mellitus" with positive certainty was considered as a different concept unit from "diabetes mellitus" with negative certainty.

Normalization of concept unit frequency: Because there was a significant difference of the average number of clinical notes between the case and the control group, we normalized the frequency of each concept unit before feeding them into a machine learning algorithm. For a patient's concept unit, the normalized frequency was calculated as: $normalized\ frequency = \frac{\sum_{i=1}^{n} frequency\ in\ clinicl\ note\ i}{n}$, where n is his/her total number of clinical notes in 2007. For example, a patient may have a total of eight clinical notes in 2007. In these eight notes, a concept unit, diabetes mellitus (SNOMED CT Concept ID: 73211009) with positive certainty, has been identified 100 times. The normalized frequency of this concept unit is then 12.5.

Machine Learning Algorithm and Evaluation: The concept units and their normalized frequencies were input as features and values into the SVM algorithm to build a classification model using Weka[14]. For both the case and the control groups, precision, recall, and F-score were calculated to evaluate the approach's performance. Precision is defined as the fraction of identified subjects that are true positive, whereas recall measures the fraction of target subjects that are identified. An F-score balances the effects of precision and recall by calculating their harmonic mean, which is defined as $F_{score} = 2 \times \frac{presision \times recall}{precison + recall}$. In order to obtain accurate estimates, we used the tenfold cross-validation method and repeated the process ten times for each test. The average and the standard deviation of each test's results are reported.

Investigation of Semantic Benefits: We designed two experiments, semantic type group and node collapse, to discover if the SNOMED CT semantic knowledge helps select features and alleviates the computational burden.

The idea of semantic types is used by Unified Medical Language System (UMLS). It aims to provide consistent categorizations of all concepts represented in all biomedical terminologies. There are currently 135 semantic type groups of UMLS. In this experiment, the identified concept units were classified into different semantic type groups, e.g. *Disease or Syndrome, Finding,* and *Sign or Symptom.* For each of these groups, the precision, recall, and F-score were calculated.

The node collapse experiment was designed to determine the effect of various SNOMED CT semantic granularities on the performance of patient identification. Semantic granularity is the degree of specific or particular detail in the definition of a group of concepts. SNOMED CT adopts a hierarchical structure, which is ordered from bottom to top and from a finest level of granularity to a coarsest level of a granularity. For example, *viral pneumonia* IS-A *Infectious pneumonia* IS-A *Pneumonia* IS-A *Lung disease*. All SNOMED CT concepts eventually are organized into nineteen top hierarchy concepts. The nineteen hierarchy concepts were called the first levels in this study. The identified branch node concepts were collapsed into various upper levels. For each level, the precision, recall, and F-score were calculated.

Results:

A total number of 29,569 distinct SNOMED CT concept units were identified from the 3,200 patients' clinical notes (a total number of 57,707). The performance of using all identified concept units as features are shown in table 1. Our approach achieved an F-score of 0.956 and 0.957 for the case and the control group respectively. For the case identification purpose, this approach produced a precision of 0.968.

	Precision	Recall	F-Score
Case	0.968±0.001	0.943±0.001	0.956±0.001
Control	0.945±0.001	0.969±0.001	0.957±0.001

Table 1: Performance by using all concept units as features.

The numbers of the identified concept units of different semantic type groups are shown in table 2. Approximately 1/5 of the identified concept units (6,457) belonged to *Disease or Syndrome* group. The other major groups were *Finding; Body Part, Organ, or Organ Component*; *Therapeutic or Preventive Procedure*; and *Sign or Symptom,* which contained 4,191, 3,644, 3,190, and 2,191 identified concept units respectively. Using only concept units of *Disease or Syndrome* group as features achieved a high performance (F-Score above 0.950) for both the case and the control groups. This indicated that this semantic type group contains the most important information for the T2DM patient identification. Using only *Finding* concept units gave a high precision (0.918) for the case group. However, the recall was relatively low (0.637). Using concept units from *Sign or Symptom* did not perform well. This may because different diabetes patients may show different signs or symptoms. This may also because the signs or symptoms of diabetes, e.g. thirst, blurred vision, dry skin, or fatigue, are not distinct from that of other diseases.

The results of the node collapse to different levels are shown in table 3. When the concept nodes collapsed to the 4th level, the number of concept units reduced to approximately one-fifth of the number of total identified concept units. The F-scores were 0.945 and 0.946 for the case and the control group respectively. Using concepts of finer levels increased the number of features; however, it did not help significantly improve the identification performance. Our results imply that the coarse level concepts, e.g. the 4th level or 5th level concepts, contain enough information for this T2DM patient identification.

Semantic Type	number of concept units	Case			Control		
		Precision	Recall	F-Score	Precision	Recall	F-Score
Disease or Syndrome	6457	0.969±0.002	0.935±0.001	0.952±0.001	0.937±0.001	0.970±0.002	0.953±0.001
Finding	4191	0.918±0.001	0.637±0.002	0.752±0.001	0.722±0.001	0.944±0.001	0.818±0.001
Body Part, Organ, or Organ Component	3644	0.730±0.003	0.580±0.003	0.646±0.002	0.652±0.002	0.786±0.003	0.712±0.002
Therapeutic or Preventive Procedure	3190	0.845±0.002	0.466±0.002	0.601±0.002	0.632±0.001	0.915±0.002	0.747±0.001
Sign or Symptom	2191	0.682±0.007	0.376±0.004	0.484±0.004	0.569±0.002	0.825±0.006	0.674±0.003
Neoplastic Process	1452	0.598±0.004	0.366±0.014	0.454±0.012	0.543±0.003	0.754±0.007	0.631±0.002
Pathologic Function	1185	0.703±0.006	0.487±0.008	0.575±0.004	0.607±0.002	0.795±0.009	0.688±0.003
Diagnostic Procedure	1114	0.584±0.002	0.813±0.008	0.680±0.003	0.693±0.006	0.422±0.008	0.524±0.005
Injury or Poisoning	1036	0.719±0.011	0.257±0.016	0.379±0.017	0.548±0.003	0.899±0.011	0.681±0.003
Body Location or Region	838	0.682±0.001	0.706±0.002	0.694±0.001	0.696±0.001	0.671±0.002	0.683±0.001
Mental or Behavioral Dysfunction	731	0.562±0.000	0.595±0.001	0.578±0.001	0.570±0.001	0.537±0.001	0.553±0.001
others	4061	0.861±0.003	0.601±0.002	0.708±0.002	0.693±0.001	0.903±0.002	0.784±0.001

Table 2: The performance of using various semantic type concept units.

Level	number of concept units	Case			Control		
		Precision	Recall	F-Score	Precision	Recall	F-Score
1	30	0.665±0.000	0.708±0.002	0.686±0.001	0.688±0.001	0.643±0.001	0.665±0.001
2	218	0.688±0.001	0.665±0.001	0.676±0.001	0.675±0.001	0.699±0.002	0.687±0.001
3	1467	0.843±0.001	0.820±0.001	0.831±0.001	0.825±0.002	0.847±0.001	0.836±0.001
4	6582	0.955±0.001	0.936±0.001	0.945±0.001	0.937±0.001	0.956±0.001	0.946±0.001
5	13116	0.958±0.001	0.938±0.001	0.948±0.001	0.939±0.001	0.959±0.001	0.949±0.001
6	21180	0.968±0.001	0.940±0.001	0.954±0.001	0.942±0.000	0.969±0.001	0.955±0.001
7	27461	0.966±0.002	0.945±0.001	0.956±0.001	0.946±0.000	0.967±0.001	0.957±0.001
8	31307	0.967±0.001	0.945±0.001	0.956±0.001	0.946±0.001	0.968±0.001	0.957±0.000
9	32785	0.967±0.001	0.944±0.001	0.956±0.001	0.946±0.001	0.967±0.004	0.956±0.002
10	33084	0.970±0.001	0.939±0.000	0.954±0.000	0.941±0.000	0.971±0.001	0.956±0.000
11	32662	0.969±0.000	0.944±0.001	0.956±0.001	0.945±0.001	0.969±0.000	0.957±0.001
12	31953	0.969±0.001	0.943±0.001	0.956±0.001	0.945±0.001	0.970±0.001	0.957±0.001

Table 3: Results of node collapse to various levels.

Level	number of concept units	Case			Control		
		Precision	Recall	F-Score	Precision	Recall	F-Score
1	30	0.665±0.000	0.708±0.002	0.686±0.001	0.688±0.001	0.643±0.001	0.665±0.001
2	218	0.688±0.001	0.665±0.002	0.676±0.001	0.675±0.001	0.699±0.002	0.687±0.001
3	1467	0.775±0.002	0.666±0.001	0.716±0.001	0.707±0.001	0.806±0.002	0.753±0.002
4	6563	0.807±0.002	0.724±0.002	0.763±0.001	0.750±0.001	0.827±0.002	0.787±0.001
5	13086	0.838±0.002	0.757±0.002	0.795±0.002	0.778±0.002	0.853±0.002	0.814±0.002
6	21130	0.854±0.002	0.773±0.002	0.811±0.002	0.793±0.002	0.867±0.002	0.828±0.002
7	27401	0.901±0.002	0.767±0.002	0.828±0.002	0.797±0.002	0.916±0.002	0.852±0.001
8	31244	0.924±0.002	0.772±0.003	0.841±0.002	0.804±0.002	0.936±0.002	0.865±0.002
9	32724	0.925±0.003	0.772±0.002	0.842±0.002	0.805±0.002	0.938±0.003	0.866±0.002
10	33021	0.927±0.002	0.772±0.003	0.842±0.002	0.804±0.002	0.940±0.002	0.867±0.002
11	32601	0.928±0.003	0.769±0.002	0.841±0.001	0.803±0.001	0.941±0.003	0.866±0.001
12	31895	0.928±0.003	0.769±0.002	0.841±0.001	0.802±0.001	0.940±0.002	0.866±0.001

Table 4: Results of node collapse to various levels after removing all concepts, the fully specified names of which contain the string "diabetes".

Disorder related concepts like "Type II diabetes mellitus (SNOMED CT Concept ID: 190384004)" are often coarse level concepts. In order to investigate if the identification process is a trivial task of looking for these disorder-related concepts, we removed all concept units from the input features if their fully specified names contain the string "diabetes" and repeated the node collapse experiment. The results are shown in table 4. The performance kept stable until the concepts collapsed into the 7th level. It is not surprising that the F-score dropped an approximate 10 percent from that of previous node collapse experiments. However, the F-scores for both the case and control groups were still at an overall acceptable level—0.850. This result suggests that the removed "diabetes" disorder-related concepts are important in this identification task while other concepts also contain enough information to classify T2DM cases and controls.

Discussion:

Our work was driven by the shortage of a general automatic approach for identifying patients with a specific phenotype. The approach we proposed takes advantage of an already existing large sample of Mayo Clinic electronic clinical notes. The approach also combines NLP, ontology, and machine learning techniques. This unique combination makes the approach novel and innovative. In addition, once the patient identifier is created, a patient identification process will be fully automatic. This would potentially save time and valuable resources.

The excellent performance of this approach on T2DM clinical notes shows its potential to be used for other patient identification tasks. The proposed approach may substantially benefit the patient recruitment process of clinical research, evidence-based healthcare, and genotype and phenotype association studies. Our results also indicated that it is promising to take advantage of SNOMED CT semantic knowledge for a pre-feature selection. This may improve the efficiency of patient identification. Concept units from the *Disease or Syndrome* semantic type group play important roles in this identification task. The coarse level concepts of SNOMED CT contain enough information for T2DM patient identification.

This preliminary evaluation is limited by an arbitrary and very narrow patient domain area (T2DM). We have planned to repeat this approach on other diseases and evaluate its performance.

Acknowledgements: W-Q. W. would like to acknowledge to Mr. Vinod and Dr. Guergana for their technical support, also acknowledge funding by the BICB Traineeship Program, UMR, MN. Support in part also arose from eMERGE project, NIH U01 HG04599.

Reference:

1. Lussier YA, Liu Y. Computational approaches to phenotyping: high-throughput phenomics. Proc Am Thorac Soc 2007;4:18-25.
2. Feero WG, Guttmacher AE, Collins FS. The genome gets personal--almost. JAMA 2008;299:1351-2.
3. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care 2005;43:480-5.
4. Kern EF, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res 2006;41:564-80.
5. Schmiedeskamp M, Harpe S, Polk R, Oinonen M, Pakyz A. Use of International Classification of Diseases, Ninth Revision, Clinical Modification codes and medication use data to identify nosocomial Clostridium difficile infection. Infect Control Hosp Epidemiol 2009;30:1070-6.
6. Chute CG. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. Proc AMIA Symp 2002:165-9.
7. Shortliffe EH, Cimino JJ. Biomedical informatics : computer applications in health care and biomedicine. 3rd ed. New York, NY: Springer; 2006.
8. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42:760-72.
9. cTAKEs. 2010. (Accessed 02/10/2010, 2010, at https://cabig-kc.nci.nih.gov/Vocab/KC/index.php.)
10. Bursi F, Weston SA, Redfield MM, et al. Systolic and diastolic heart failure in the community. JAMA 2006;296:2209-16.
11. Pakhomov SS, Hemingway H, Weston SA, Jacobsen SJ, Rodeheffer R, Roger VL. Epidemiology of angina pectoris: role of natural language processing of the medical record. Am Heart J 2007;153:666-73.
12. Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997.
13. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Boston: Pearson Addison Wesley; 2006.
14. Witten IH, Frank E. Data mining : practical machine learning tools and techniques. 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman; 2005.
15. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7:298-303.
16. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006;81:741-8.
17. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc 1997;4:484-500.
18. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak 2008;8 Suppl 1:S2.
19. EMERGE project. (Accessed 07/31/2009, at https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page.)