Is the inter-patient coincidence of a subclinical disorder related to EHR similarity?

Lawrence W.C. Chan, Iris F.F. Benzie
Department of Health Technology and Informatics
Hong Kong Polytechnic University
Hong Kong, China
wing.chi.chan@inet.polyu.edu.hk
iris.benzie@inet.polyu.edu.hk

Y. Liu
Department of Mechanical Engineering
National University of Singapore, Singapore
mpeliuy@nus.edu.sg

C.R. Shyu
Informatics Institute
University of Missouri
Columbia, MO 65211-2060, USA
ShyuC@missouri.edu

Abstract—**Electronic Health Record (EHR) provide clinical evidence for identifying subclinical diseases and supporting decisions on early intervention. Simple string matching cannot link up the conceptually similar but verbally different clinical terms in patient records, limiting the usefulness of EHR. A novel ontological similarity matching approach supported by the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is proposed in this paper. The disease terms of a patient record are transformed into a vector space so that each patient record can be characterized by a feature vector. The similarity between the new record and an existing database record was quantified by a kernel function of their feature vectors. The matches are ranked by their similarity scores. To evaluate the proposed matching approach, medical history and carotid ultrasonic imaging finding were collected from 47 subjects in Hong Kong. The dataset formed 1081 pairs of patient records and the ROC analysis was used to evaluate and compare the accuracy of the ontological similarity matching and the simple string matching against the presence or absence of carotid plaques identified in ultrasound examination. It was found that the simple string matching randomly rated the record pairs but the ontological similarity matching provided non-random rating.**

Keywords-SNOMED; similarity; clinical decision support; Electronic Health Record

I. INTRODUCTION

ElectroNIC Health Record (EHR) system is comprised of computer software and hardware components for providing the archiving and communications of patient-centered clinical information throughout the episodes of the care of each patient. In Hong Kong, the EHR system of the Hong Kong Hospital Authority (HKHA) is one of the world's largest integrated longitudinal EHR systems [1]. The general use of EHR focuses on the longitudinal study of the clinical history of the individual patient only.

A. Clinical Decision Support

To support the health care professionals to make clinical decisions and maintain quality of care, clinically meaningful search for the similar patient records becomes a important feature of HER system. Simple string matching has long been used to search for patient records exactly or partially matching with the given keywords. However, a large portion of conceptually match records are most likely missing in the search results. For example, "Coronary Artery Disease" and "Myocardial Ischemia" are closely related in medical concept but the simple string matching cannot link these two disease terms and also the patient records containing them. To address this issue, medical ontology could be considered and incorporated into the search algorithm.

B. Medical Ontology

Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) has been widely adopted as a standard for formulating medical concepts. Over 361800 unique concepts with 975000 descriptions have been covered in SNOMED-CT as of 2004 [2, 3]. SNOMED-CT defines semantic relationships including an extensive "is-a" and "inverse-is-a" structures. Through these defined relationships, the relative closeness between two concepts in a record is measured by "edge counting" the semantic distance along the connecting path in the ontological hierarchy [2, 4-7]. The edge counting method has been applied to PubMed document clustering and the performance was comparable to the alternative methods [7], such as information content based measure associating the probabilities with concepts in the ontology [5, 8].

C. EHR Similarity

Based on a well-established ontology, the vector (space) model offers simple parallel evaluation of similarity between queries and documents through the construction of the feature vectors, in which every term at a particular level of the considered ontology is weighted with a real number in [0,1].

The weight is zero if feature term or its descendent is not present in the query/documents, otherwise, with positive number reflecting the relative importance in the query/documents [9, 10]. The similarity between the given query and every document can be calculated and the search algorithm could return a list of documents above the similarity threshold. When the query and document are two patient records, such EHR similarity is referred to as ontological similarity in this paper.

Simple string matching is a common search algorithm bundled with many EHR systems. A list of hits will be generated according to the provided search keys. The hits

include those exact matched items and some similar items in wordings, if wild card function and logical operators are used. However, it is not able to map a query to the similar health records in the database with respect to the medical concepts because the simple string matching cannot identify the association of two terms that are completely different in wording but conceptually related (e.g. heart disease and cardiac arrhythmia). A combination of string-based and tag (concept)-based retrieval yielded 95% mean recall of relevant records in contrast to 54% using string-based model [11].

D. Research Question and Significance

In this paper, we proposed an ontological similarity matching algorithm based on SNOMED-CT and tested its performance against the patient linkage with the subclinical carotid plaques using the receiver-operating characteristic (ROC) analysis [12]. We examined whether the EHR similarity based on medical ontology is related to the inter-patient coincidence of subclinical disorder. For comparison, this hypothesis will be also tested for simple string matching.

We hypothesized the EHR similarity based on simple string matching is a random rater because there is a lack of clinical information or coding about the diabetic complications in the health records. On the other hand, medical ontology could relate the medical concepts like hypertension and atherosclerosis and has good potential to detect undocumented subclinical disorders from the patient records. The findings of this study could be used as a reference for the development of the case-based CDS for diabetes management.

II. METHODS

A. SNOMED

Medical ontology, SNOMED-CT, was used in this research work. The research team of this study have been offered an affiliate license of UMLS (license code: 23044A180) for academic use since 1 July 2007. It is assumed that the disease terms extracted from the patient records are translated and indexed with the SNOMED-CT terms, and referred to as "EHR terms" in this paper. The ontological vector model comprises a set of feature terms at fixed level of the "is-a" hierarchy, thus forming a feature space. When level 4 is considered as the pre-defined feature space, the feature term set includes "Disorder of body system", "Disorder of soft tissue", etc, as shown in figure 1.

B. Similarity Measure

Let f_i, m, d_j and n be the i^{th} feature term, the number of terms in the feature space, the j^{th} EHR term and the number of terms in the EHR respectively. The semantic distance between terms f_i and d_j is denoted by $s_{ij} \in [0,\infty]$ and its value is generated by the following if-then rules.

(A) If d_j is descendant of f_i in the SNOMED-CT hierarchy, then s_{ij} is the number of "is-a" links from d_j to f_i.
(B) If d_j is not descendant of f_i, then $s_{ij} = \infty$.
(C) If d_j is the same as f_i, then $s_{ij} = 0$.

The feature vector, v, is a linear array of m+102 elements given by the following expression.

$$v = [a_1\ a_2\ \ldots\ a_m\ g_1\ g_2\ e_1\ e_2\ \ldots\ e_{100}]^T \quad (1)$$

Where g_1 and g_2 are Boolean variables for female and male respectively; e_k is Boolean variable for age = k $\in$ [1, 99] and age $\geq$ 100 for k =100; a_i is the weight of feature term f_i in characterizing a patient record, denoted by $\{d_j\ \forall j \in [0,n]\}$. The element a_i is computed using the following formula.

$$a_i = \frac{1}{1 + \min_{j=1\ldots n} s_{ij}} \quad (2)$$

Similar to the path-length measure [3], the value of a_i quantifies the projection of a patient record into a particular feature term. According to the above-mentioned rule (B) and the equation (2), those feature terms irrelevant to the patient record are weighed with 0 in the feature vector. Therefore, a large number of zeros are found in the feature vector and the non-zeros form a minority of vector elements making the patient record distinguishable from the others. Gender and age are also included in the feature vector because they also contribute to the diabetic complications.

The similarity measure between the query and an existing patient record is denoted by sim (Q, D). Direction cosine and Euclidean distance are two types of most commonly used similarity measures. Direction cosine is a kernel function of the feature vectors [9], given by the following formula.

$$sim(Q,D) = \frac{Q \bullet D}{|Q||D|} \quad (3)$$

where Q and D represent the feature vectors of the query and an existing patient record respectively; $\bullet$ is the inner product between two vectors; and |.| is the geometric vector length.

The shortest distance between two vectors represents the Euclidean distance, given by the following formula.

$$d = |Q - D| \quad (4)$$

The distance is zero if both Q and D have same length and direction. Otherwise, it returns a positive value, giving the geometrical distance between Q and D. The similarity measure based on Euclidean distance is given by the following formula.

$$sim(Q,D) = \frac{1}{1 + |Q - D|^2} \quad (5)$$

C. Validation

To test the proposed algorithm, a dataset was collected from 47 subjects in Hong Kong. All were Type II diabetic patients, aged 46 to 60 years, non-smokers, and with no record of stroke or coronary heart disease. Human ethics approval and

informed consent were obtained before data collection. Each subject completed a questionnaire, which collected information including age, gender and medical history, and an ultrasound examination of the carotid arteries of each subject was also performed. A radiologist confirmed the presence or absence of carotid plaque for each subject. This confirmed manifestation of vascular abnormality was regarded as the reference match between any two subjects. The diseases shown in the medical history became the EHR terms and then the feature vector of each subject was computed based on the SNOMED-CT "is-a" hierarchy. There were totally 1081 pairs of feature vectors. Between each pair of patient records, a similarity score was obtained. For simple string matching, the "matching score" is simply given by the number of co-occurring terms in a pair of patient records. Through the ROC analysis, the proposed algorithm and the simple string matching were compared with respect to the accuracy in matching subjects with or without the carotid plaque. The pairs were sorted in ascending order of the score of interest. A threshold value between the minimum and the maximum values of the score was set amid every two consecutive sorted scores to generate the true positive rate (sensitivity) and the false positive rate (1-specificity) based on the ranking collated with the sorted reference matches. The Receiver-Operating Characteristic (ROC) curve was drawn by plotting the true positive rate against the false positive rate for all the threshold values. The empirical area under the ROC curve (AUROC) was estimated using trapezoidal rule. Wilcoxon statistic was exploited to estimate the standard error (SE) of AUROC, indicating the sampling variability for the null hypothesis "true area=0.5" [12]. The asymptotic 95% confidence interval (CI) was given by [AUROC-1.96*SE, AUROC+1.96*SE]. The performance of the ontological vector model and the simple string matching was tested against the criterion of a non-random rater of record pairs that the empirical AUROC was significantly different from 0.5.

III. Results

The similarity and matching scores of 1081 EHR pairs were checked against the reference values of similarity. The scores were sorted in ascending order and the true positive and the false positive rates were obtained for each threshold. After the construction of the ROC curves, the AUROCs and the corresponding asymptotic 95% CIs were calculated and shown in table I. The results showed that the exact matching was a random rater in evaluating the similarity in terms of carotid plaque presence as the AUROC is not significantly different from 0.5. The proposed vector model at any considered ontological level performed as a non-random rater in term of carotid plaque identification as the AUROC was significantly greater than 0.5. The ontological vector model at the SNOMOD-CT level 4 yielded the highest accuracy of plaque identification where the AUROCs were 0.578 and 0.587 for direction cosine and Euclidean distance respectively.

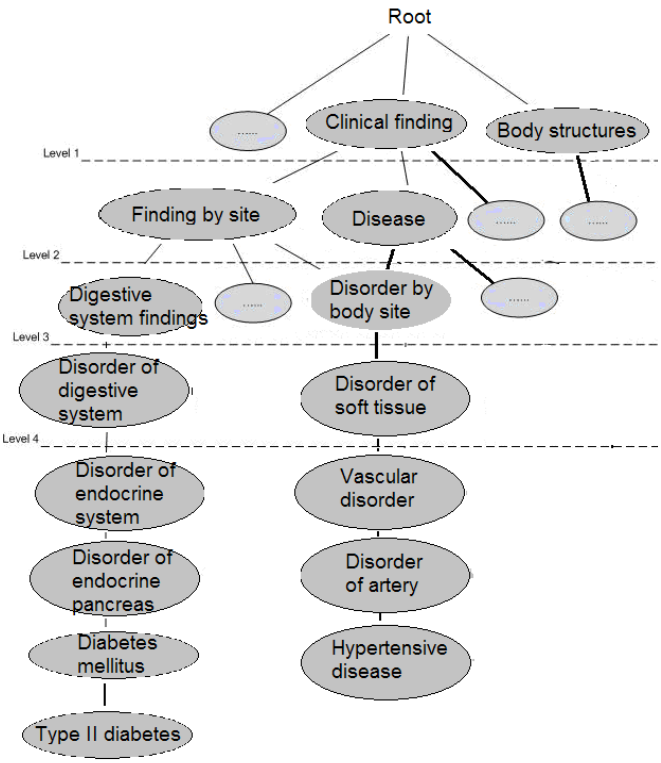

Figure 1. Part of the "is-a" relationships between terms in SNOMED-CT hierarchy.

Query patient record	Existing patient record
Patient Q	Patient A
Gender:female/Age:46	Gender:female/Age:61
SNOMED-CT:	SNOMED-CT:
Diabetes Mellitus, Non-Insulin-Dependent (C0011860)	Diabetes Mellitus, Non-Insulin-Dependent (C0011860)
Dental caries (C0011334)	Hypertensive disease (C0020538)
Gingivitis (C0017574)	Hyperostosis (C0020492)
Bronchitis (C0006277)	Frozen shoulder (C0311223)
Chest Pain (C0008031)	Iritis (C0022081)

Figure 2. Example of two patient records.

IV. Discussion

The findings of ROC analysis revealed that the simple string matching paired up patient records by chance, but the ontological similarity matching gave a non-random rater with respect to the links with carotid plaques. Figure 2 provides an example of the two subjects from the dataset explaining the difference between two approaches.

It was found that only two items, "Diabetes Mellitus, Non-Insulin-Dependent" and "Female" were exactly matched when comparing patients Q and A. The matching score, equal to 2, was relatively low. However, carotid plaques were confirmed in these two subjects and the matching score of 2 was not high

enough to rank this pair. On the other hand, the ontological similarity matching provided a more precise feature extraction.

Although the terms "Dental caries", "Gingivitis" and "Bronchitis" of patient Q were not verbally matched with "Hypertensive disease", "Frozen shoulder" and "Iritis" of patient A, these terms activate the feature terms "Finding of head and neck region", "Disorder of head", "Disorder of soft tissue" and "Inflammation of specific body structure or tissue" at the SNOMED-CT level 4 and contributed considerable portion of the similarity score. The relatedness of these features to the atherosclerosis at the carotid arteries explained why the similarity score can rate the likelihood of carotid plaque presence. The relatedness is valid due to the evidence provided by the clinical and epidemiological studies that inflammation and its markers play a very important role in the pathogenetic mechanism in atherosclerosis [13]. Moreover, in the medical concept, the carotid arteries at the neck region are responsible for the blood supply to the soft tissue of the head region. Therefore, the proposed algorithm demonstrated the potential in detecting subclinical manifestation of complications, which have been not diagnosed in the medical history of the query record but can be identified through the similarity search in the existing EHR database.

The accuracy of carotid plaque identification was the highest for the ontological similarity matching at the SNOMED-level 4 because level 4 has the highest degree of granularity when comparing with levels 1-3. A study of the inter-patient distance metrics using SNOMED-CT claimed that the usefulness of ontology principles ("is-a" relationships) as tools for a particular purpose, which is the identification of carotid plaques in this work, is highly dependent of the quality and granularity of the terminology [2]. Therefore, the choice of an adequate high level as the feature space is essential for developing a promising similarity measure.

This study explored the potential of the ontological similarity measure as a rater for comparing the query and database patient records. The experiment was performed under the assumption that the disease terms of both health records can be extracted and mapped to unique concepts in medical ontology. In the scenario of CDS, the EHR similarity can be used to query health records similar to that of the patient of interest in the consultation and the clinicians can use the retrieved as references for their decision making.

TABLE I. PERFORMANCE OF THE ONTOLOGICAL SIMILARITY MATCHING AT DIFFERENT SNOMED-CT LEVEL AND THE SIMPLE STRING MATCHING ACCORDING TO THE ROC ANALYSIS

Algorithm	Level	AUROC	95% CI
Ontological similarity matching: Direction cosine	1	0.542	(0.507, 0.577)
	2	0.535	(0.500, 0.569)
	3	0.572	(0.538, 0.606)
	4	0.578	(0.544, 0.612)
Ontological similarity matching: Euclidean distance	1	0.563	(0.528, 0.597)
	2	0.565	(0.530, 0.599)
	3	0.583	(0.549, 0.617)
	4	0.587	(0.553, 0.622)
Simple string matching	-	0.474	(0.439, 0.509)

V. CONCLUSION

The findings showed that the degree of EHR similarity based on medical ontology between two health records is associated with the agreement of subclinical atherosclerosis between two patients. Case-based CDS could be further developed based on such association and applied to diabetes management and prevention of cardiovascular diseases.

REFERENCES

[1] Cheung, N.T., Fung, V., Wong, W.N., Tong, A., Sek, A., Greyling, A., Tse, N., Fung, H., 2007. Principles-based medical informatics for success--how Hong Kong built one of the world's largest integrated longitudinal electronic patient records. Studies in health technology and informatics 129 (1), 307-310.

[2] Melton, G.B., Parsons, S., Morrison, F.P., Rothschild, A.S., Markatou, M., and Hripcsak, G., 2006. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 39 (6), 697-705.

[3] Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G., 2007. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40 (3), 288-299.

[4] Knappe, R., Bulskov, H., Andreasen, T., 2007. Perspectives on Ontology-based Querying. Int J of Intell Syst 22, 739-761.

[5] Petrakis, E.G.M., Varelas, G., Hliaoutakis, A., Raftopoulou, P., 2006. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. In: the proceedings of the 4th Workshop on Multimedia Semantics (WMS'06), pp. 44-52.

[6] Slimani, T., Yaghlane, B.B., Mellouli, K., 2006. A New Similarity Measure based on Edge Counting. In: Proceedings of World Academy of Science, Engineering and Technology, pp. 34-38.

[7] Zhang, X., Jing, L., Hu, X., Ng, M., Zhou, X., 2007. A comparative study of ontology based term similarity measures on PubMed document clustering. In: Lecture Notes in Computer Science 4443, pp. 115-126.

[8] Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A., 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19 (10), 1275–1283.

[9] Salton, G., 1991. Developments in Automatic Text Retrieval. Science 253 (5023), 974-980.

[10] Salton, G., Buckley, C., 1991. Global Test Matching for Information Retrieval. Science 253 (5023), 1012-1015.

[11] Mikkelsen, G., Aasly, J., 2002. Manual semantic tagging to improve access to information in narrative electronic medical records. Int J Med Inform 65 (1), 17–29.

[12] Hanley, J.A., McNeil, B.J., 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143, 29-36.

[13] Pearson, T.A., Mensah, G.A., Alexander, R.W., Anderson, J.L., Cannon, R.O., Criqui, M., Fadl, Y.Y., Fortmann, S.P., Hong, Y., Myers, G.L., Rifai, N., Smith, S.C., Taubert, K., Tracy, R.P., Vinicor, F., 2003. Markers of Inflammation and Cardiovascular Disease - Application to Clinical and Public Health Practice: A Statement for Healthcare Professionals From the Centers for Disease Control and Prevention and the American Heart Association. Circulation 107, 499-511.