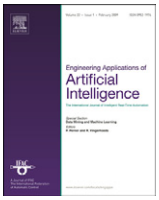
Engineering Applications of Artificial Intelligence 24 (2011) 1398–1409

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



A SNOMED supported ontological vector model for subclinical disorder detection using EHR similarity

L.W.C. Chan [a],*, Y. Liu [b], C.R. Shyu [c], I.F.F. Benzie [a]

[a] *Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong*
[b] *Department of Mechanical Engineering, National University of Singapore, Singapore 117576, Singapore*
[c] *MU Informatics Institute, The University of Missouri, USA*

ARTICLE INFO

Available online 15 June 2011

Keywords:
Electronic Health Record
Vector model
SNOMED
Similarity
Clinical decision support
Atherosclerosis

ABSTRACT

Electronic Health Records (EHR) form a valuable resource in the healthcare enterprise because clinical evidence can be provided to identify potential complications and support decisions on early intervention. Simple string matching, the common search algorithm, is not able to map a query to the similar health records in the database with respect to the medical concepts. A novel ontological vector model supported by the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) is proposed in this paper to project the disease terms of a health record to a feature space so that each health record can be characterized using a feature vector, giving a fingerprint of the record. The similarity between the query and database health records was measured by similarity measures of their feature vectors and string matching score respectively. Three types of similarity measures were considered in this study, namely, Euclidean distance (ED), direction cosine (DC) and modified direction cosine (mDC). Medical history and carotid ultrasonic imaging findings were collected from 47 subjects in Hong Kong. The dataset formed 1081 pairs of health records and ROC analysis was used to evaluate and compare the accuracy of the ontological vector model and simple string matching against the agreement of the presence or absence of carotid plaques identified by carotid ultrasound between two subjects. It was found that the score generated by simple string matching was a random rater but the ontological vector model was not. In other words, the degree of health record similarity based on the ontological vector model is associated with the agreement of atherosclerosis between two patients. The vector model using feature terms at the SNOMED-CT level 4 gave the best performance. The performance of mDC was very close to that of ED and DC but the properties of mDC make it more suitable for the retrieval of similar health records. It was also shown that the ontological vector model was enhanced by the support vector classifier approach.

1. Introduction

Electronic Health Record (EHR) system is patient-centered information resource supported by computer software and hardware infrastructure. In the clinical workflow, EHR system facilitates the archiving and communications of clinical information of each patient throughout the episodes of care. More than a data repository of digitized health records, EHR system provides core functionalities in eight categories, including electronic health information/data capture, order entry/order management, results management, administrative processes, electronic connectivity, clinical decision support (CDS), health outcomes reporting and patient support (Institute of Medicine and Board on Health Care Services, 2003). Among these eight categories, clinical decision support enhances clinical performance by adopting computer software applications to generate information items, such as drug alerts, other rule-based alerts, reminders, clinical guidelines and pathways. With the clinical decision support (CDS) tools, EHR system can effectively support the patient care and reduce the cost, time and errors in clinical workflow (Morris, 2005). Since the idea of CDS has first emerged, a lot of research endeavors have dedicated to the research and development of tools, methods and approaches for pattern discovery and predictive analysis. Some of those research studies explored the associations between diagnoses based on datasets of collected health records. IBM's HealthMiner and Molecular Concept Map (MCM) are two examples of data mining tools, which have been successfully applied in two different research studies to identify the related diagnoses among the health records (Mullins et al., 2005; Hanauer et al., 2009). Another research study investigates a method for selecting

* Corresponding author. Tel.: +852 34008596; fax: +852 23624365.
E-mail address: wing.chi.chan@inet.polyu.edu.hk (L.W.C. Chan).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.05.013

the threshold of Chi-square test in analyzing co-occurring features, such as disease, finding and medication, of a large dataset of discharge summaries (Cao et al., 2007). These tools are very important and useful for the clinical experts to gain insights of disease patterns and linkages from the large EHR database. Besides CDS, clinical trials require the database retrieval of relevant cases for subject recruitment. A concept-based document search engine was implemented with an EHR system and evaluated through a retrospective clinical-epidemiological study targeting syphilis cases (Schulz et al., 2008). A research study proposed a sequence alignment strategy for finding patients with similar treatment histories in the temporal order of drugs given to the patients (Lee and Das, 2010). The approach could be useful for treatment cohort identification in clinical trials.

1.1. Motivation

As the methods for exploring disease patterns and linkages have been well developed, identifying similar health records based on similarity in disease pattern becomes a novel topic and growing research area in CDS. In some stages of clinical decision making processes, uncertainty exists and it is difficult to model the relevant medical domain knowledge in logical representations like rules. The similar health records retrieved using similarity matching algorithm are more concrete and convincing for medical experts to express their domain knowledge on particular disease patterns than the rule-based representation (Kong et al., 2008). For this reason, case-based reasoning approach presents similar health records to the clinicians as suggested solutions in many CDS applications, e.g. stress diagnosis based on finger temperature signals. To achieve the desired matching, the biomedical features were appropriately selected for those similarity matching algorithms based on the a-priori medical knowledge. For example, age, gender, room temperature, hours since meal, food and drink taken and finger temperature were considered as features for matching patients on similar stress level. Better goodness-of-fit can be attained when the parametric similarity measure is learnt through the adjustment of weights by domain experts or using the empirical data (Begum et al., 2009).

For matching patients with the same subclinical disorder like atherosclerosis, the etiology and underlying mechanism have not been completely known and clearly explained. The feature selection is not feasible as there are only a few variables that have significant association with the subclinical disorder and the association is usually very weak. An example of such variable is C-reactive protein (CRP). It was demonstrated in recent research studies that the elevation of CRP indicates increased risk of atherothrombotic events. However, it was shown in a large cross-sectional study that CRP is weakly associated with the manifestations of prevalent atherosclerosis, such as the intima-media thickness and ankle/brachial blood pressure index (Folsom et al., 2001). As all the available features are considered, a large dataset is required to train the parametric similarity measure. Otherwise, the trained measure may not be so accurate to match similar patient records. Therefore, non-parametric similarity measures are considered in this study.

Many classic similarity measures are non-parametric, including Euclidean distance (ED) and direction cosine (DC) (Qian et al., 2004). ED takes into the account the lengths of two vectors and the angle between them but DC depends on the deviation of vector direction only. Two similar health records retrieved based on DC could be very different in content. On the other hand, ED involves higher computational load than DC because ED considers all the vector components but DC needs not compute the zero-valued vector components, which are very often observed in the information retrieval applications. Combining the properties of ED and DC may yield a similarity measure addressing both precision and computational load issues. Further, the application of supervised machine learning approach could enhance the performance of these similarity measures through the relative weights of feature terms estimated by the training dataset.

1.2. Research questions

In this study, all the available disease terms in the health records are considered as the features and non-parametric similarity measure is used because most of the current EHR search engines are generally used for searching health records subject to any possible desired clinical terms and these tools were developed based on simple string matching of clinical terms. Thus, it is intuitive to hypothesize that a non-parametric similarity measure is a random scoring system in the agreement of a subclinical disorder between two patients. To cope with the randomness, the semantic relationships defined by medical ontology may help to estimate the closeness between patients of interest. Such kind of closeness, referred to as ‘‘clinical distance’’, was estimated by the semantic distance of the ‘‘minimal-cost’’ path in the ontology (Melton et al., 2006). As medical ontology can be used to align the synonymous and related clinical terms to unique concepts and form the feature space for the similarity measure, the likelihood of a common subclinical disorder found in two patients may be reflected by the similarity score based on medical ontology. To test the hypotheses for simple string matching and medical ontology, this study is aimed to examine whether the degree of similarity between two health records is associated with the subclinical disorder agreement between two patients, for the similarity measures based on simple string matching and medical ontology respectively. Further, a modified version of DC, referred to as mDC, is proposed in this paper and its performance in ranking the subclinical disorder agreement between two patients will be compared with that of ED and DC. This study is also aimed to examine whether the mDC outperforms the DC and ED in the estimation of patient similarity in terms of atherosclerosis. The implementation of the ontological model using machine learning approach will be also evaluated against the identification of subclinical disorder in the test dataset.

1.3. Core contribution and novelty of the work

Atherosclerosis is a major cause of cardiovascular disease (CVD). The American Heart Association (AHA) reported that no previous symptoms were observed in 50% of men and 64% of women who died suddenly of coronary heart disease (CHD) (Lloyd-Jones et al., 2009). Therefore, there is a need for search engine, which can retrieve similar cases from EHR system to provide concrete evidence for assessing the CVD risk of the asymptomatic individuals of interest. Once the above-mentioned research question is answered, it will bring out the potential of a medical-ontology-based search algorithm for identifying patients with similar CVD risk which is unlikely achieved by the current approach of simple string matching. From this study, the ontology-specific feature space and the derived similarity measure thereof will lead to a search engine, which will act as a simple, non-invasive and inexpensive tool to present similar cases to clinicians during consultation of asymptomatic patients. If the similar cases are already of CVD, the cases could be the evidence for predicting the future of the patients of interest and justifying the early intervention and prevention.

As a diabetic complication, atherosclerosis and its related information are not commonly documented in clinical practice. A study in southwest London showed that all of 17 participating general practices input the codes of cardiovascular risk factors, including urine protein, blood pressure, serum cholesterol, dietary advice given and diabetes mellitus, in the health records, but only one of them input the code of Framingham cardiac risk score (the chance of developing cardiovascular disease within a certain number of years) and three and nine of them input the codes of diabetic retinopathy and neuropathy respectively (Gray et al., 2003). Thus, there is a lack of clinical information or coding about the diabetic complications in the health records and it spotlights the limitation of simple string matching in the retrieval of cases with diabetic complications. On the other hand, medical ontology has good potential for linking related concepts like hypertension and atherosclerosis and is superior to simple string matching in digging undocumented subclinical disorders from EHR system. The proposed search engine and findings of this study could be used as a reference for the future development of the case-based CDS for diabetes management.

Ontology vector models have been widely applied to more information retrieval applications. Two classic similarity measures, ED and DC, have their own pro and con in terms of the computational load and precision. This paper suggests a new similarity measure, mDC, which conserves the advantageous properties of ED and DC. Fast and precise retrieval can be achieved by using mDC.

2. Background and related work

2.1. Local and worldwide context of EHR use

In Hong Kong, the EHR system of the Hong Kong Hospital Authority (HKHA), called electronic Patient Record (ePR), is one of the world's largest integrated longitudinal EHR (Cheung et al., 2007). The common use of EHR focuses on the longitudinal study of the clinical history of the concerned patient. The exploration of the similarity or difference between patients may be useful for providing useful information for clinical decision support. The inter-patient similarity measures assess the degree of closeness between patients or cases in a database so as to discover analogous cases, such as those patients who qualify for an experimental oncology chemotherapy protocol (Melton et al., 2006). Except for health policy purposes, it is rare to find research studies looking into the patterns and associations among the patient records. The EHR database is still an under-explored valuable resource that has archived a variety of cases with complete recovery, good prognosis, medical errors or ineffective treatments. Despite the popularity of simple string matching, EHR search engine finding the conceptual linkage between patients would play a more important role in supporting patient care, because the retrieved records are not only the demographic data of particular patients but also refer to concrete concepts such as "their parts, diseases, therapies and lesions", which can be used as references for the health care professionals to make diagnosis and plan treatment (Ceusters and Smith, 2006). The superior performance of the search algorithm based on conceptual linkage has been demonstrated in the information retrieval over a large document repository on the World Wide Web (Castells et al., 2007).

The success of search engines, like Google, in searching World Wide Web contents sheds light on the potential of search engines for EHR. Researcher developed the Electronic Medical Record Search Engine (EMERSE) for retrieving health records for research and data abstraction (Hanauer, 2006). The contribution of a search engine is to provide a rapid method for patient history review with retrospective chart so that the events of interest could be easily noted.

2.2. Medical terminologies and ontologies

Nowadays, more systematic medical terminologies and standard codes are provided to the health care professionals during their input, update and review of the patient records so that the data contained in these records become more objective and consistent, with less variability in terminologies and data entry habit among professionals. In Sweden, general practitioners use the International Classification of Disease (ICD) as a tradition for diagnostic labeling in the EHR (Nilsson et al., 2000) and a large proportion of problems are coded in general practice in Stockholm where the completeness (number of codes per problem=0.9) and correctness (97.4%) of diagnostic codes are high (Nilsson et al., 2003). A study in Norway revealed that the mean precision of record retrieval, i.e. the proportion relevant to a particular concept of all retrieved EHR elements, was 100% for manual semantic tagging and 37% for string matching (Mikkelsen and Aasly, 2002). In the United Kingdom, general practitioners in primary care currently code clinical problems using Read codes developed in the 1980s to ensure the construction of accurate disease registries (Gray et al., 2003). Among medical ontologies, Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) has been widely adopted as a standard for formulating medical concepts. In the 2010 international release of SNOMED-CT, more than 291,000 active concepts are included and the descriptions of these concepts are over 758,000. The concepts are organized into hierarchies, in which 823,000 relationships enable the consistency of data retrieval and analysis. The breadth of SNOMED-CT terms and concepts were evaluated in a study through the clinicians' coding of diagnosis and problem lists within a computerized physician order entry (CPOE) system in the United States, resulting in 98.5% concept coverage (Wasserman and Wang, 2003). Another study examined whether the clinical information retrieval at an electronic medical record data warehouse is simplified by leveraging the knowledge built into the hierarchical structure of SNOMED-CT and the results showed that the recall of SNOMED-based record retrieval, i.e. the proportion retrieved of the EHR elements relevant to a particular concept, was greater than 89% (Lieberman et al., 2003). Besides the classification of terms, SNOMED-CT incorporates concepts of entities on the ontological view. The entities include universals, kinds, attributes or properties in the world (Smith, 2006). With linkages between entities, concepts provide more accurate and substantial representation of entities than the codes and semantic tags only. However, there exists an ontological issue from the perspective of SNOMED-CT that the same entity, e.g. colon polyp, can result in different codes on successive clinical visits, e.g. intestinal polyp, polyp and malignant neoplasm of colon (Ceusters and Smith, 2006). Though some search engines may facilitate wild card function and logical operators to formulate the query statement, 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 (Mikkelsen and Aasly, 2002).

2.3. Semantic relationships using SNOMED-CT

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 (Knappe et al., 2007; Melton et al., 2006; Petrakis et al., 2006; Slimani et al., 2006; Zhang et al., 2007) and the clinical distance, the closeness between patients of interest is obtained by taking the semantic distance of the "minimal-cost" path (Melton et al., 2006). The path-length measure is a mathematical representation of semantic measure defined as the inverse of the node number on the shortest path between two concepts and the adaptation by Leacock and Chodorow (1998) further scales the shortest path length by the maximum depth attained by the concepts of interest in the hierarchy (Pedersen et al., 2007; Knappe et al., 2007; Zhang et al., 2007). The edge counting method has been applied to PubMed document clustering and the performance was comparable to the alternative methods (Zhang et al., 2007), such as information content based measure associating the probabilities with concepts in the ontology (Lord et al., 2003; Petrakis et al., 2006).

2.4. Sub-ontology extraction for specialization area

Ontologies can be regarded as a meta-data source of terms, covering the domains of interest. In the medical applications, the analysis of medical data is commonly specialty-specific and the usage of a medical ontology is limited to a subset of the entire meta-data. A sub-ontology view extraction approach was proposed for semantic profiling for particular specialties and the process is referred to as Materialized Ontology View Extraction (MOVE) (Bhatt et al., 2009). The application of MOVE contextualizes the user's specialization area on the types of terms (concepts) so that the semantic range is significantly reduced to form the sub-ontology or partial views of the ontology of interest.

For the measurement of the similarity between two health records using medical ontology, the occurring terms or concepts will be very few or none if the semantic range is narrowed. As the application of MOVE for matching patients with atherosclerosis could limit the sub-ontology to the taxonomy branches like cardiovascular disease and hypertensive disease, the extracted sub-ontology may not cover some related concepts like disorder of soft tissue and inflammatory disease. The calculated similarity score may not reflect the agreement of atherosclerosis between two patients.

2.5. Application of gene ontology for searching biomedical literature

PubMed is a tremendous database of the abstracts of biomedical articles annotated with the terms from the Medical Subject Headings (MeSH) controlled vocabulary. To respond to a query, PubMed search engine matches the query term with the articles' MeSH terms and their descendants in the hierarchy and returns all the matches to the user interface. Currently, over 20 million articles are indexed by PubMed and a query may return hundreds or even thousands of matches. With the support of PubMed search interface, the matches can be sorted by first/last author, title and publication date/venue. However, some relevant articles might be published a few years ago and their author names, titles and publication venues may not reflect the articles' relevancy. The relevant or desired articles may be moved to the end of the query result by PubMed search interface, limiting its usefulness in literature search. To provide a better solution, GoPubMed is a web-based search engine, which relays query keywords to PubMed, extracts gene ontology (GO) terms from the query results of the matched abstracts and allows users to browse the induced ontology for desired abstracts (Doms and Schroeder, 2005). The induced ontology is a minimal subset of GO derived from the GO terms extracted from the retrieved abstracts. The tool provides the users with a query-specific taxonomy for further specifying the query, instead of a list of abstracts ranked by the frequencies of the terms occurring in PubMed. GoPubMed demonstrated that the documents deeply hidden by the PubMed-sorted matches can be explored easily through the application of medical ontology. This statement is also applicable to the retrieval of health records similar to a query record. There may be hundreds or even thousands of health records matched with a query record according to the associated disease terms, especially in hospital clusters or regional health enterprises. If the matched health records are sorted by the patient name, age, clinician name or admission date, the agreement of subclinical disorder between the patients of the top matches and the query may be less likely observed. Ranking the health record pairs by similarity measure based on medical ontology could feature particularly those patients with the same subclinical disorder.

2.6. Vector model

With a well-established term set, such as medical ontology, the vector (space) model offers simple parallel evaluation of similarity between queries and documents through the construction of the query and document vectors, in which every single term in the term set is weighted with zero if is not present in the query/documents, otherwise, with positive number reflecting the relative importance in the query/documents (Salton and Buckley, 1991; Salton, 1991). Given a query of weighted terms for identifying similar documents in a repository indexed with weights of the same term set, the inner product between the normalized elements of the query and document vectors ranks the query-document similarity and returns those documents above the similarity threshold in decreasing order of ranks. The update of query and document vectors is easy for query reformulation, document changes and other purposes due to the intuitive vector mathematics. The similarity can be quickly computed because there is most likely a substantial number of zeros in the query and document vectors, like "Boolean retrieval", eliminating a large amount of numerical operations. Such classic vector model has been applied to the information retrieval over a large document repository on the World Wide Web (Castells et al., 2007). The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm was used to compute the weights of the vector automatically. The similarity measure between the query and a document is computed as the direction cosine of their respective vectors. It was shown that a combination of ontology and keyword search gave the best performance in terms of precision versus recall for 20 queries to a corpus of documents from CNN Website.

2.7. Similarity measures

Once the ontological vectors are obtained, the closeness between two patients can be quantified by a score generated by a mathematical function of the two vectors, called similarity measure. There are many different choices of similarity measures and the appropriate choice is of considerable importance to applications, such as pattern classification, clustering and information retrieval (Cha, 2007). Euclidean distance (ED) and direction cosine (DC) are two similarity measures commonly used for ranking relatedness in the vector model due to the simplicity in computation. These two measures achieved very similar level of recalls and precisions in content-based image retrieval (Qian et al., 2004). ED is grounded on the fact that the shortest geometric distance between two points is a straight line. ED compares both the lengths and directions of two vectors. The comparison is very stringent because the two patients are conceptually the closest if and only if the ontological vectors are the same, i.e. the patients have the same diseases. DC was widely adopted in automatic text retrieval as it ranks the similarity according to the angle between two vectors only (Salton, 1991). The patients with different diseases could be ranked with the maximum of DC, i.e. unity, if the ontological vectors are pointing at the same direction. DC is still very popular because it consumes less computation resource than ED due to the fact that the zero vector elements are not involved in the calculation of DC while ED takes into the account all the vector elements.

3. Material and methods

In this study, similarity measure is regarded as a rater for comparing the query health record and a database health record. It is assumed that both health records contain disease terms that can be mapped to unique concepts in medical ontology. The database health records are already stored in the data repository of EHR system. The query health record belongs to the patient of interest in the consultation and based on which the clinician requests for similar health records as references for supporting the clinical decision making. Fig. 1 shows the examples of the query health record (query EHR in step 1) and a retrieved database health record thereof (retrieved EHR in step 3) and illustrates how a search algorithm ranks the database health records through the scores generated by a similarity measure (health record list sorted in the descending order of similarity score).

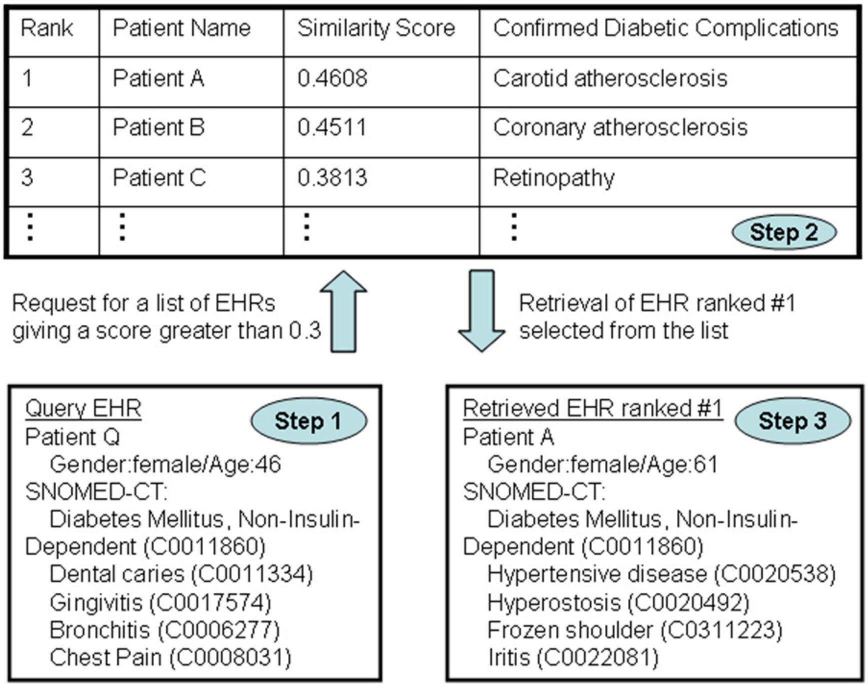
3.1. Ontology vector model

SNOMED-CT, the ontology used in this research work, was developed by the College of American Pathologists and now constitutes part of the Unified Medical Language System established by the National Library of Medicine (NLM) (Pedersen et al., 2007). The researchers of this study have been offered an affiliate license of UMLS (license code: 23044A180) for academic use since 1 July 2007. Tree structure of SNOMED-CT hierarchy is illustrated in Fig. 2, demonstrating the "is-a" relationships between terms extending from root to two sample diseases, type II diabetes and cardiac arrhythmia. The terms of SNOMED-CT are organized into levels and connected with "is-a" links. Root is at level 0 and the terms connected to root are at level 1. The "is-a" links directionally map terms from one level to the next level based on the semantic relationships. No "is-a" links can be drawn from one level to the same or previous level. For example, there is a "is-a" link from "Disorder of digestive system" to "Disorder of pancreas" as disorder of pancreas is subsumed into the disorder of digestive system, however, there should not be any "is-a" link between terms at level 4, say "Disorder of digestive system" and "Finding of trunk structure".

As the natural language processing is not a focus of this work, the free text disease terms were converted to SNOMED concepts manually. Therefore, the health records, irrespective of the type of database records or queries, are indexed with the SNOMED-CT terms, referred to as 'EHR terms' in this paper. The proposed ontology vector model considers a set of feature terms at fixed level of the "is-a" hierarchy, which makes up a feature space. For example, the feature term set consists of "Disorder of digestive system", "Finding of trunk structure", etc., if level 4 is considered as the pre-defined feature space. Let f_i, m, d_j and n be the ith feature term, the number of terms in the feature space, the jth EHR term and the number of terms in the health record, respectively. The semantic distance between f_i and d_j is denoted by $s_{ij} \in [0, \infty]$ and the value is measured according to the following rules:

(A) If d_j is the 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 the descendant of f_i, then $s_{ij} = \infty$.
(C) If d_j is the same as f_i, then $s_{ij} = 0$.

Consider the term "Type II diabetes" as an example and level 4 terms as the feature space. As illustrated in Fig. 2, the semantic distance between "Type II diabetes" and the feature term "Disorder of digestive system" is 4 because there are four "is-a" links



Rank	Patient Name	Similarity Score	Confirmed Diabetic Complications
1	Patient A	0.4608	Carotid atherosclerosis
2	Patient B	0.4511	Coronary atherosclerosis
3	Patient C	0.3813	Retinopathy
⋮	⋮	⋮	⋮

Step 2
Request for a list of EHRs giving a score greater than 0.3
Retrieval of EHR ranked #1 selected from the list
Step 1
Step 3

Query EHR
Patient Q
Gender:female/Age:46
SNOMED-CT:
Diabetes Mellitus, Non-Insulin-Dependent (C0011860)
Dental caries (C0011334)
Gingivitis (C0017574)
Bronchitis (C0006277)
Chest Pain (C0008031)

Retrieved EHR ranked #1
Patient A
Gender:female/Age:61
SNOMED-CT:
Diabetes Mellitus, Non-Insulin-Dependent (C0011860)
Hypertensive disease (C0020538)
Hyperostosis (C0020492)
Frozen shoulder (C0311223)
Iritis (C0022081)

Fig. 1. Demonstration of the query/retrieve function of the EHR search engine with the ontological vector model: in step 1, the query EHR is provided and its feature vector is produced. The feature vectors of query EHR and each EHR in the database are compared using similarity score. In step 2, a list of EHRs with score greater a threshold, say 0.3, shows up in decreasing order of similarity and allows the users to pick one to retrieve. In step 3, the selected EHR is displayed. The codes inside the brackets are concept IDs of SNOMED-CT.

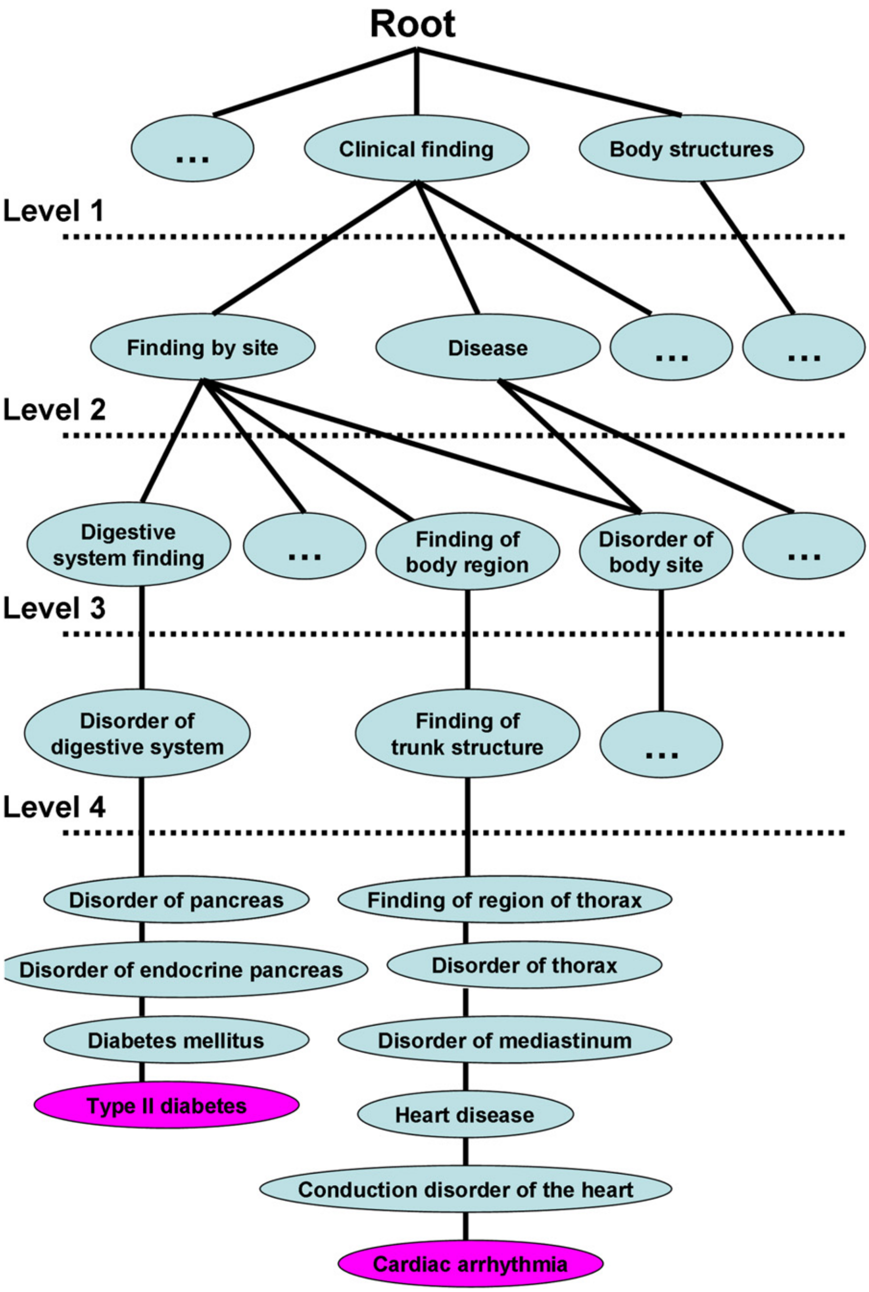


Fig. 2. Part of the "is-a" relationships between terms in SNOMED-CT hierarchy.

from Type II diabetes to "Disorder of digestive system". The semantic distance between "Type II diabetes" and the feature term "Finding of trunk structure" is infinity because "Type II diabetes" is not the descendant of "Finding of trunk structure". Note that this is just an example to illustrate the calculation of semantic distance between two concepts and it does not mean that "Type II diabetes" is not related to "Cardiac arrhythmia". In fact, there is another branch expanded from "Finding of trunk structure", along which the leaf "Type II diabetes" can be reached. That branch is not shown in Fig. 2 for simplicity. The feature vector, v, is a linear array of $m+102$ elements given by the following expression:

$$v=[a_1a_2\ldots a_mg_1g_2e_1e_2\ldots e_{100}]^T \tag{1}$$

where g_1 and g_2 are Boolean variables for female and male respectively; e_k is Boolean variable for age$=k\in[1,99]$, except e_{100} for age≥ 100; a_i is the weight of term f_i with respect to an EHR, represented by $\{d_j\ \forall j\in[0,n]\}$. The element a_i is computed using the following formula:

$$a_i=\frac{1}{1+\min\limits_{j=1\ldots n} s_{ij}} \tag{2}$$

The value of a_i, quite similar to the path-length measure (Pedersen et al., 2007), projects the health record onto a feature term. It weighs those relevant feature terms with positive values and those irrelevant feature terms with zeros in the feature vector. Thus the feature vector comprises of a vast majority of zeros and a small number of non-zeros, providing a discriminative fingerprint of the health record. Gender and age are included in the feature vector because they are also contributing factors of many complications.

3.2. Similarity measures: ED, DC, mDC

Through the alignment of EHR terms to the medical ontology, an ontological vector can be generated to each patient. It is natural to hypothesize that those diabetic patients having similar disease profile (and thus similar ontological vectors) will be subjected to similar risk of cardiovascular diseases. Therefore, the similarity measure of ontological vector could link similar patients together with respect to the degree of cardiovascular risk. For ontological vector model, three types of similarity measures are considered in this study. The detailed descriptions are given as follows.

3.2.1. Euclidean distance (ED)

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

$$d_{ED} = |Q-D| \tag{3}$$

where Q and D represent the ontological vectors of two health records; The distance is zero if both vectors are identical. Otherwise, it gives a positive value, increasing with the geometrical distance between two vectors. The similarity measure based on ED is given by the following formula:

$$sim_{ED}(Q,D) = \frac{1}{1+|Q-D|^2} \tag{4}$$

3.2.2. Direction cosine (DC)

The similarity score between the two ontological vectors, denoted by, $sim_{DC}(Q,D)$, is a kernel function computed as the direction cosine between their feature vectors (Salton, 1991), given by the following formula:

$$sim_{DC}(Q,D) = \frac{Q \cdot D}{|Q||D|} \tag{5}$$

where Q and D represent two ontological vectors; $\cdot$ is the inner product between two vectors; and $|.|$ is the geometric vector length.

3.2.3. Modified direction cosine (mDC)

The similarity score between the two ontological vectors, denoted by, $sim_{mDC}(Q,D)$, is a kernel function computed as the direction cosine between their augmented feature vectors, given by the following formula:

$$sim_{mDC}(Q,D) = \frac{(Q,1)\cdot(D,1)}{|Q,1||D,1|} \tag{6}$$

where Q and D represent two ontological vectors; $\cdot$ is the inner product between two vectors; and $|.|$ is the geometric vector length. The feature vectors are augmented by appending a unity element.

3.2.4. Properties of mDC

Unlike DC, mDC does not have the problem of numerical overflow when Q or (and) D is (are) zero vector(s). If Q is zero vector, mDC can be simplified by the following formula:

$$sim_{mDC}(\vec{0},D) = \frac{1}{|D,1|} \tag{7}$$

It is easy to observe that the value of (7) is always nonnegative. If neither Q nor D are zero vectors, Eq. (6) can be rewritten by the following formula:

$$sim_{mDC}(Q,D) = \frac{\cos\theta + (1/|Q||D|)}{\sqrt{(1+(1/|Q|^2))(1+(1/|D|^2))}} \tag{8}$$

where θ is the angle between Q and D, always lying within $[0,\pi/2]$ as the vector elements are nonnegative. Thus, mDC is always nonnegative. According to (8), mDC is equal to zero when θ is equal to $\pi/2$, the length of one vector is positive and the other's length tends to infinity. This agrees with the property of ED under the condition that two points are very far apart in the coordinate system.

When two feature vectors Q and D are equal, their lengths are equal and the angle θ is 0. The value of mDC reaches 1, the maximum of Eq. (8), if and only if the condition, $Q=D$, is satisfied. ED also attains its maximum if and only if the same condition is satisfied. This property makes mDC perform better than DC and similar to ED.

3.3. Scoring and search algorithm for CDS

The query EHR can be compared with every single EHR in the database according to this similarity score and the retrieval output can be selected subject to a threshold of score and ranked with the score values. Fig. 1 shows a flow chart demonstrating the conceptual interactive search engine empowered by the proposed ontological vector model.

The list of the relevant health records provided in step 2 of Fig. 1 indicates the confirmed diabetic complications. In this example, the most similar patient (with rank #1) is confirmed to have carotid atherosclerosis and thus the patient Q (with the query EHR) may have plaques at the carotid arteries with a chance reflected by the score 0.4608. Carotid ultrasound examination could be suggested to the patient Q though the score is not too high. Further, the similarity is expected to be valued around the middle of the range (i.e. 0.5), leading to a narrow range of similarity values. It is because no two patient records are exactly the same or completely different, especially for those in the same EHR database or from the same medical consultation center.

3.4. Data collection and analysis

To test the hypothesis, a dataset was collected from 47 subjects in Hong Kong. All were Type II diabetic patients, aged 46–62 years, non-smokers and with no record of stroke or coronary heart disease. Human ethics approval and informed consent were obtained before data collection. This study did not retrieve any health records of these subjects from EHR databases because most of the subjects were consulting doctors at different private clinics and it was difficult to obtain their health records directly due to the resistance from the private clinics. Instead of direct retrieval, data items that are commonly stored in health records were collected from the subjects through questionnaire survey. Each subject completed a questionnaire, which collected information including age, gender and medical history with disease terms. A research assistant graduated from medical school advised the subjects in completing the questionnaires, curated the collected data items and coded the disease terms with concept IDs of SNOMED-CT. 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 value to verify the similarity between any two subjects. The diseases shown in the medical history became the EHR terms and then the feature vector of each subject (health record) was computed based on the SNOMED-CT "is-a" hierarchy. Based on such taxonomy, a feature vector was generated for each health record. Totally, 47 feature vectors were generated. The similarity measure is a pairwise rater for comparing any two health records (any two feature vectors). The number of pairwise similarity scores is therefore given by $N(N-1)/2=47(47-1)/2=1081$ for all pairwise combinations of 47 subjects. There were totally 1081 pairs of feature vectors (health records). Between each pair of health records, a similarity score was obtained using the similarity measure. For simple string matching, the "matching score" is simply given by the number of co-occurring terms in a pair of health records. For the query EHR and the retrieved EHR shown in Fig. 1, only the concept "Diabetes Mellitus, Non-insulin-Dependent" and gender "Female" co-occur in both health records and thus the similarity score is 2 using simple string matching. Note that the similarity score using simple string matching is not bounded by one and its value cannot be directly compared with that using ontological vector model.

Through the Receiver-Operating Characteristic (ROC) analysis, the ontological vector model using ED, DC and mDC, and the simple string matching were compared with respect to the accuracy in matching subjects with or without the carotid plaque based on their scoring approaches. The health record pairs for all subjects were sorted in ascending order of the similarity scores. 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 value of similarity. The 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" (Hanley and McNeil, 1982). 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 that the empirical AUROC was significantly different from 0.5, implying that the model is not randomly rating the health record pairs.

3.5. Ontological model implementation using support vector approach

Once the proposed ontological model is proved to be a non-random rater of health record pairs, it can be implemented to identify patients potentially with subclinical disorder. Direct application of the above-mentioned similarity measures may not yield good performance for case retrieval because the relative importance of each feature term cannot be reflected by the non-parametric dot product or Euclidean distance. Supervised machine learning approach was applied to implement the ontological model. support vector (SV) classifier would be a suitable parametric model to accommodate the a-priori weighting subject to the relevancy of the feature terms (Vapnik, 1995). The SV approach makes use of the Lagrange multipliers, nonnegative real numbers, and a training dataset is required to obtain the values of the Lagrange multipliers through the constrained minimization. The resultant model of the SV classifier is a linear regression function of the pre-defined kernels plus a bias. The kernels are, for the proposed vector model, simply the dot product of the feature vectors of a health record in the training dataset and the health record of interest, satisfying the Mercer's condition (Chan et al., 2001). The classifier $F(.)$ is given by the following equation:

$$F(Q)=\sum_{i=1}^{N}(\overline{\alpha}_i-\underline{\alpha}_i)K(Q,D_i)+b \tag{9}$$

where N is the size of training dataset; $\overline{\alpha}_i$ and $\underline{\alpha}_i$ are the Lagrange multipliers and b is the bias; D_i is the feature vector of the ith health record in the training dataset; $K(Q,D)$ could be a kernel defined by the following expression:

$$K(Q,D)=Q\cdot D \tag{10}$$

If Q and D are feature vectors already normalized with their vector lengths, this kernel is equivalent to the similarity score used in this study. It is easy to observe that the classifier function $F(.)$ is simply a linear combination of the similarity score between the feature vectors of each health record in the training dataset, D, and the health record of interest, Q. The classifier will give a positive (negative) value to indicate the health record of interest is positive (negative) case according to the knowledge learnt from the training dataset. By replacing the kernels in the classifier with the dot product of feature vectors, we can obtain an alternative representation of the classifier as given below:

$$F(Q)=Q\cdot\left(\sum_{i=1}^{N}(\overline{\alpha}_i-\underline{\alpha}_i)D_i\right)+b \tag{11}$$

From this representation, each feature term is associated with a weight, specifying the relevancy of the feature term to the reference similarity value of interest according to the training dataset. This set of a-priori weights is formed by the vector elements of the following linear combination of Lagrange multipliers and feature vectors of health records in training dataset

$$\sum_{i=1}^{N}(\overline{\alpha}_i-\underline{\alpha}_i)D_i \tag{12}$$

The support vector classifier could be nonlinear when we consider kernels, such as polynomial, B-spline and radial basis functions (RBF), instead of linear kernel, i.e. dot product. In this study, Bioinformatics toolbox of Matlab® was used to construct the SV classifier. With randomization, feature vectors and the corresponding atherscolerosis labels of 32 subjects form the training dataset and the rest form the test dataset. After the SV classifier was trained, its performance on the test dataset was evaluated by the accuracy, precision and recall.

4. Experimental results

Levels 1–4 were considered individually as the feature space. Of 47 health records, the feature vectors were computed. Considering level 4 as the feature space and two health records mentioned in Fig. 1, Tables 1 and 2 illustrate how the EHR terms are projected to the feature terms through the semantic distance given by Eq. (2) and the calculation of their feature vector elements (except those elements for gender and age). It was found that patients Q and A have, respectively, 22 and 23 non-zero feature vector elements only but there are totally 6964 feature terms at level 4. In the calculation of similarity score between them, we need to consider the co-occurring non-zero elements, which involved 12 feature terms only. Table 3 shows the alignment of the two feature vectors (including the elements for gender and age as well). The data obtained from these two subjects yielded a similarity score 0.4608 after taking the direction cosine of the two feature vectors. Both subjects had carotid plaques as confirmed by ultrasound examination and thus this pair of patients (health records) was determined as relevant and labeled by a dichotomous value 1 as the reference value of similarity. If both subjects had no carotid plaques, the pair was also determined as relevant and labeled by a dichotomous value 1 as the reference value of similarity. If either of the subjects had carotid plaques, the pair was determined as irrelevant and labeled by a dichotomous value 0 as the reference value of similarity.

Table 1
Calculation of feature vector elements (except for gender and age) of patient Q with SNOMED-CT level 4 as the feature space: the second to sixth columns show the number of "is-a" links from the feature terms to the EHR terms. Note that blank cells represent infinity and the feature terms with zero weights (at the last row) are not shown here for simplicity.

Feature term at level 4 of SNOMED-CT	Concept ID of Patient Q's EHR terms					Feature vector element
	C0011860	C0011334	C0017574	C0006277	C0008031	
Disorder of digestive system	4	5	3			1/(1+3)=0.25
Disorder of carbohydrate metabolism	3					1/(1+3)=0.25
Abdominal organ finding	5					1/(1+5)=0.17
Finding of pancreas	4					1/(1+4)=0.2
Finding of trunk structure	6			4	2	1/(1+2)=0.33
Disorder of body system	4	6	4	4		1/(1+4)=0.2
Disorder of trunk	5			3		1/(1+3)=0.25
Disorder of body cavity	4	3	5	2		1/(1+2)=0.33
Bacterial infectious disease		4				1/(1+4)=0.2
Mouth and/or pharynx finding		5	5			1/(1+5)=0.17
Finding of head and neck region		6	5			1/(1+5)=0.17
Infection by site		4				1/(1+4)=0.2
Injury of anatomical site		1				1/(1+1)=0.5
Disorder of head		4	3			1/(1+3)=0.25
Disorder of soft tissue			2			1/(1+2)=0.33
Gingivae finding			2			1/(1+2)=0.33
Inflammation of specific body structures or tissue			2	2		1/(1+2)=0.33
Disorder of respiratory system				3		1/(1+3)=0.25
Lower respiratory tract finding				2		1/(1+2)=0.33
[D]Symptoms, signs and ill-defined conditions					3	1/(1+3)=0.25
Pain/sensation finding					4	1/(1+4)=0.2
Pain of truncal structure					1	1/(1+1)=0.5
…						0

Table 2
Calculation of feature vector elements (except for gender and age) of patient A with SNOMED-CT level 4 as the feature space: the second to sixth columns show the number of "is-a" links from the feature terms to the EHR terms. Note that blank cells represent infinity and the feature terms with zero weights (at the last row) are not shown here for simplicity.

Feature term at level 4 of SNOMED-CT	Concept ID of Patient A's EHR terms					Feature vector element
	C0011860	C0020538	C0020492	C0311223	C0022081	
Disorder of digestive system	4					1/(1+4)=0.2
Disorder of carbohydrate metabolism	3					1/(1+3)=0.25
Abdominal organ finding	5					1/(1+5)=0.17
Finding of pancreas	4					1/(1+4)=0.2
Finding of trunk structure	6					1/(1+6)=0.14
Disorder of body system	4	4	4	4	5	1/(1+4)=0.2
Disorder of trunk	5					1/(1+5)=0.17
Disorder of body cavity	4					1/(1+4)=0.2
Disorder of cardiovascular system		3				1/(1+3)=0.25
Disorder of soft tissue		3				1/(1+3)=0.25
Blood vessel finding		3				1/(1+3)=0.25
Disorder of connective tissue			3			1/(1+3)=0.25
Disorder of musculoskeletal system			3	3		1/(1+3)=0.25
Growth alteration			2			1/(1+2)=0.33
Bone finding			2			1/(1+2)=0.33
Joint finding				3		1/(1+3)=0.25
Finding of limb structure				3		1/(1+3)=0.25
Inflammation of specific body structures or tissue				2	4	1/(1+2)=0.33
Disorder of extremity				2		1/(1+2)=0.33
Visual system disorder					5	1/(1+5)=0.17
Globe finding					4	1/(1+4)=0.2
Finding of head and neck region					6	1/(1+6)=0.14
Disorder of head					5	1/(1+5)=0.17
…						0

The similarity and matching scores of 1081 health record 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 4. 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 demonstrated a non-random rater of health record pairs 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 for ranking the agreement of carotid plaque identification in subject pairs. The AUROCs are 0.578, 0.587 and 0.584 for DC, ED and mDC, respectively.

Table 3
Alignment of feature vectors of patients Q and A. Note that the feature terms with zero weights in both EHRs are not shown here for simplicity.

Feature term at level 4 of SNOMED-CT	Patient Q	Patient A
Disorder of digestive system	1/(1+3)=0.25	1/(1+4)=0.2
Disorder of carbohydrate metabolism	1/(1+3)=0.25	1/(1+3)=0.25
Abdominal organ finding	1/(1+5)=0.17	1/(1+5)=0.17
Finding of pancreas	1/(1+4)=0.2	1/(1+4)=0.2
Finding of trunk structure	1/(1+2)=0.33	1/(1+6)=0.14
Disorder of body system	1/(1+4)=0.2	1/(1+4)=0.2
Disorder of trunk	1/(1+3)=0.25	1/(1+5)=0.17
Disorder of body cavity	1/(1+2)=0.33	1/(1+4)=0.2
Bacterial infectious disease	1/(1+4)=0.2	0
Mouth and/or pharynx finding	1/(1+5)=0.17	0
Finding of head and neck region	1/(1+5)=0.17	1/(1+6)=0.14
Infection by site	1/(1+4)=0.2	0
Injury of anatomical site	1/(1+1)=0.5	0
Disorder of head	1/(1+3)=0.25	1/(1+5)=0.17
Disorder of soft tissue	1/(1+2)=0.33	1/(1+3)=0.25
Gingivae finding	1/(1+2)=0.33	0
Inflammation of specific body structures or tissue	1/(1+2)=0.33	1/(1+2)=0.33
Disorder of respiratory system	1/(1+3)=0.25	0
Lower respiratory tract finding	1/(1+2)=0.33	0
[D]Symptoms, signs and ill-defined conditions	1/(1+3)=0.25	0
Pain/sensation finding	1/(1+4)=0.2	0
Pain of truncal structure	1/(1+1)=0.5	0
Disorder of cardiovascular system	0	1/(1+3)=0.25
Blood vessel finding	0	1/(1+3)=0.25
Disorder of connective tissue	0	1/(1+3)=0.25
Disorder of musculoskeletal system	0	1/(1+3)=0.25
Growth alteration	0	1/(1+2)=0.33
Bone finding	0	1/(1+2)=0.33
Joint finding	0	1/(1+3)=0.25
Finding of limb structure	0	1/(1+3)=0.25
Disorder of extremity	0	1/(1+2)=0.33
Visual system disorder	0	1/(1+5)=0.17
Globe finding	0	1/(1+4)=0.2
…	0	0
Female	1	1
Male	0	0
Aged_46	1	0
Aged_61	0	1

Table 4
Performance of the ontological vector model at different SNOMED-CT level and the simple string matching according to the ROC analysis.

Algorithm	**SNOMED-CT level for feature space**	**AUROC**	**Asymptotic 95% CI**
Ontological vector model 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 vector model 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)
Ontological vector model Modified direction cosine	1	0.562	(0.528, 0.597)
	2	0.564	(0.530, 0.599)
	3	0.579	(0.545, 0.613)
	4	0.584	(0.550, 0.619)
Simple string matching	–	0.474	(0.439, 0.509)

The above findings demonstrated the possibility for applying the ontological model to identify patients with the same subclinical disorder. The SV classifiers with linear kernel, RBF kernel and 3rd order polynomial kernel were trained and tested. It was found that both classifiers with linear and RBF kernels gave 0% precision and 0% recall. On the other hand, the classifier with 3rd order polynomial kernel yielded 50% precision and 25% recall and the overall accuracy was 73.3%, which is much higher than 58.4%, the AUROC attained by mDC. The performance justifies the implementation of the ontological model by SV classifier.

5. Discussion

5.1. Implications of findings

The findings of ROC analysis revealed that the simple string matching paired up ‘‘similar’’ health records by chance, but the ontological vector model acted as a non-random rater with respect to the agreement of carotid plaque identification between two subjects. The example of the two subjects shown in Fig. 1, explained 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 similarity score generated by simple string matching is equal to 2. The value was relatively low as carotid plaques were confirmed in these two subjects. The unmatched EHR terms, such as ‘‘Hypertensive disease’’ and ‘‘Chest pain’’, are also related to diabetic complications but the simple string matching cannot reflect the importance of these terms. On the other hand, the ontological vector model provided a more complete fingerprinting of the health record, as illustrated in Tables 1–3. 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 triggers 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 generated by ontological vector model can rate the likelihood of carotid atherosclerosis occurring in a subject pairs. Such relatedness can be explained by the findings of the clinical and epidemiological studies that inflammation and its markers play a very important role in the pathogenetic mechanism in atherosclerosis (Pearson et al., 2003). 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. That's why the feature terms "Finding of head and neck region" and "Disorder of head" should be involved in the calculation of similarity score. Therefore, the proposed ontological vector model demonstrated the potential in uncovering subclinical manifestation of complications, which have been not diagnosed or documented in the medical history but can be identified through the similar health records retrieved from the existing EHR database.

The accuracy of carotid plaque identification was the highest for the ontological vector model at the SNOMED-level 4 because level 4 has higher degree of granularity than levels 1–3. A study of the inter-patient distance metrics using SNOMED-CT claimed that the usefulness of ontology principles as tools for a particular purpose, is highly dependent of the quality and granularity of the terminology (Melton et al., 2006). This claim is further justified in this study. Therefore, the choice of level for the feature space is essential for developing a promising similarity measure.

Among the three similarity measures, the performance was very close. ED gave the highest AUROC, mDC was at the middle and DC gave the lowest AUROC at any level. It showed the potential of mDC for saving the computational load and at the same time providing better performance than the classic DC. Further, the SV machine learning implementation of the ontological model was further justified by the highly improved performance in identifying subjects with atherosclerosis.

5.2. Suggested enhancement of ontological vector model

It is not surprising to get the low AUROC for the ontological vector model because of two main reasons. First, atherosclerosis develops silently through the mediation of multiple risk factors (Dzau et al., 2002). The absence of carotid plaque in the ultrasound examination does not mean that those high-risk subjects will not build up any carotid plaque in the coming years. The reference value of similarity was adopted without the consideration of time at which the plaques were detected, increasing the number of false positives. Second, there exists limitation in this approach as it simply considers some relations such as proximity on the body. When Canonical Clinical Problem Statement System (CCPSS) is incorporated into the SNOMED-CT hierarchy, "Shoulder Injury Gunshot Wound" (C0748662) and "Frozen Shoulder" (C0311223) are both having "is-a" relationships with "Joint Problem" (C0575044). It implies that a patient had a gunshot wound to the shoulder may be linked to another patient had frozen shoulder but they are less likely having the same subclinical disorder as the causalities of the two terms are completely different. To prevent this problem, the relationships should be weighted according to their causalities. Relationships with the injury should be modulated by smaller weights when the considered subclinical disorder is atherosclerosis. Third, all the feature terms at the same level were considered but not all of them are significantly related to the carotid atherosclerosis, for example, "Infection by site" and "Lower respiratory tract finding". The *a-priori* weighting of the feature terms could benefit the classification of patients with respect to the atherosclerotic risk. The SV classifier achieved such supervised learning and the accuracy of the subclinical disorder identification was significantly raised.

6. Conclusions

This paper proposes ontological vector model for searching health records in database, which are expected to be similar to the query EHR, with respect to the carotid plaque presence or absence. The medical ontology, SNOMED-CT, was used to form the feature space for fingerprinting the health records. The results showed that the three similarity measures, ED, DC and mDC, for ontological vector model were non-random raters of health record pairs but the scores generated by the simple string matching identified the similar pairs by chance. In other words, the degree of similarity between two health records is associated with the agreement of subclinical atherosclerosis between two patients using the ontological vector model. Though the performance of three measures was very close, the modified direction cosine has several advantages over classic direction cosine and Euclidean distance in terms of numerical overflow and computational load. Further, support vector classifier could be an enhanced parametric implementation of the proposed vector model to give a weighting of the feature terms with respect to the relevancy of the complication of interest. It was found that the accuracy of carotid plaque identification was significantly improved.

Acknowledgments

This research is supported by the PolyU internal research funds "A-PH74: Artificial intelligent modeling of inter-relationships between biomarker profile and haemodynamic features of atherosclerosis in Type II Diabetes Mellitus", "A-SA33: Sketching a Multi-dimensional Arterial Health Status Map based on Fuzzy Expert System, Health Domain Expertise and Published Relative Risk Findings" and "A-PC0B: Development of a UMLS-integrated Ontology-based Similar ePR Search Engine".

References

Begum, S., Ahmed, M.U., Funk, P., Xiong, N., Schéele, B.V., 2009. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. Computational Intelligence 25 (3), 180–195.

Bhatt, M., Rahayu, J.W., Soni, S.P., Wouters, C., 2009. Ontology driven semantic profiling and retrieval in medical information systems. Journal of Web Semantics 7 (4), 317–331.

Cao, H., Hripcsak, G., Markatou, M., 2007. A statistical methodology for analyzing co-occurrence data from a large sample. Journal of Biomedical Informatics 40 (3), 343–352.

Castells, P., Fernández, M., Vallet, D., 2007. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Transactions on Knowledge Data Engineering 19 (2), 261–272.

Ceusters, W., Smith, B., 2006. Strategies for referent tracking in electronic health records. Journal of Biomedical Informatics 39 (3), 362–378.

Cha, S.H., 2007. Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences 4 (1), 300–307.

Chan, W.C., Chan, C.W., Cheung, K.C., Harris, C.J., 2001. On the modelling of nonlinear dynamic systems using support vector neural networks. Engineering Applications of Artificial Intelligence 14, 105–113.

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.

Doms, A., Schroeder, M., 2005. GoPubMed: exploring pubmed with the gene ontology. Nucleic Acids Research 33 (Web-Server-Issue), 783–786.

Dzau, V.J., Braun-Dullaeus, R.C., Sedding, D.G., 2002. Vascular proliferation and atherosclerosis: new perspectives and therapeutic strategies. Nature Medicine 8 (11), 1249–1256.

Folsom, A.R., Pankow, J.S., Tracy, R.P., Arnett, D.K., Peacock, J.M., Hong, Y., Djoussé, L., Eckfeldt, J.H., 2001. Investigators of the NHBLI Family Heart Study. Association of C-reactive protein with markers of prevalent atherosclerotic disease. American Journal of Cardiology 88 (2), 112–117.

Gray, J., Orr, D., Majeed, A., 2003. Use of Read codes in diabetes management in a south London primary care group: implications for establishing disease registers. BMJ 326, 1130–1133.

Hanauer, D.A., 2006. EMERSE: the electronic medical record search engine. In: AMIA Annual Symposium Proceedings, pp. 941.

Hanauer, D.A., Rhodes, D.R., Chinnaiyan, A.M., 2009. Exploring clinical associations using '-omics' based enrichment analyses. PLoS One 4 (4), e5203.

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.

Institute of Medicine, Board on Health Care Services, 2003. Key Capabilities of an Electronic Health Record System: Letter Report. National Academies Press, Washington, DC (Retrieved May 29, 2010, from〈http://www.prorec.it/documenti/EHR_Report_Final.pdf〉).

Knappe, R., Bulskov, H., Andreasen, T., 2007. Perspectives on ontology-based querying. International Journal of Intelligent Systems 22 (7), 739–761.

Kong, G.L., Xu, D.L., Yang, J.B., 2008. Clinical decision support systems: a review on knowledge representation and inference under uncertainties. International Journal of Computational Intelligence Systems 1, 159–167.

Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, pp. 265–283.

Lee, W.N., Das, A.K., 2010. Local alignment tool for clinical history: temporal semantic search of clinical databases. In: Proceedings of AMIA 2010 Symposium, pp.437–441.

Lieberman, M.I., Ricciardi, T.N., Masarie, F.E., Spackman, K.A., 2003. The use of SNOMED CT simplifies querying of a clinical data warehouse. In: Proceedings of the 2003 AMIA Annual Symposium, pp. 910.

Lloyd-Jones, D., Adams, R., Carnethon, M., Simone, G.E., Ferguson, T.B., Flegal, K., et al., 2009. Heart Disease and Stroke Statistics 2009 Update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 119 (3), e21–e181.

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.

Melton, G.B., Parsons, S., Morrison, F.P., Rothschild, A.S., Markatou, M., Hripcsak, G., 2006. Inter-patient distance metrics using SNOMED CT defining relationships. Journal of Biomedical Informatics 39 (6), 697–705.

Mikkelsen, G., Aasly, J., 2002. Manual semantic tagging to improve access to information in narrative electronic medical records. International Journal of Medical Informatics 65 (1), 17–29.

Morris, B., 2005. Electronic health records. Student BMJ 13, 89–132.

Mullins, I.M., Siadaty, M.S., Lyman, J., Scully, K., Garrett, C.T., Miller, W.G., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S., Knaus, W.A., 2005. Data mining and clinical data repositories: insights from a 667,000 patient data set. Computers in Biology Medicine 36 (12), 1351–1377.

Nilsson, G., Ahlfeldt, H., Strender, L.E., 2003. Textual content, health problems and diagnostic codes in electronic patient records in general practice. Scandinavian Journal of Primary Health Care 21 (1), 33–36.

Nilsson, G., Petersson, H., Ahlfeldt, H., Strender, L.E., 2000. Evaluation of three Swedish ICD-10 primary care versions: reliability and ease of use in diagnostic coding. Methods of Information in Medicine 39, 325–331.

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.

Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G., 2007. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40 (3), 288–299.

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 Fourth Workshop on Multimedia Semantics (WMS'06), pp. 44–52.

Qian, G., Sural, S., Gu, Y., Pramanik, S., 2004. Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In: Proceedings of 2004 ACM Symposium on Applied Computing, pp. 1232–1237.

Salton, G., 1991. Developments in automatic text retrieval. Science 253 (5023), 974–980.

Salton, G., Buckley, C., 1991. Global test matching for information retrieval. Science 253 (5023), 1012–1015.

Schulz, S., Daumke, P., Fischer, P., Müller, M., 2008. Evaluation of a document search engine in a clinical department system. In: Proceedings of the AMIA 2008 Symposium, pp. 647–651.

Slimani, T., Yaghlane, B.B., Mellouli, K., 2006. A new similarity measure based on edge counting. In: Proceedings of the World Academy of Science, Engineering and Technology, pp. 34–38.

Smith, B., 2006. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. Journal of Biomedical Informatics 39 (3), 288–298.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Wasserman, H., Wang, J., 2003. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. In: Proceedings of the 2003 AMIA Annual Symposium, pp. 699–703.

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

Further reading

Chi, P.-H., Pang, B., Korkin, D., Shyu, C.-R., 2009. Efficient SCOP-fold classification and retrieval using index-based protein substructure alignments. Bioinformatics 25, 2559–2565.

Häyrinen, K., Saranto, K., Nykänen, P., 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. International Journal of Medical Informatics 77 (5), 291–304.