Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 39 (2006) 697–705

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Inter-patient distance metrics using SNOMED CT defining relationships ☆

Genevieve B. Melton [a,*], Simon Parsons [b], Frances P. Morrison [c], Adam S. Rothschild [c], Marianthi Markatou [d], George Hripcsak [c]

[a] *Department of Surgery, The Johns Hopkins Medical Institutions, Baltimore, MD, USA*
[b] *Department of Computer and Information Science, Brooklyn College, Brooklyn, NY, USA*
[c] *Department of Biomedical Informatics, Columbia University, New York, NY, USA*
[d] *Department of Biostatistics, Columbia University, New York, NY, USA*

Received 10 October 2005
Available online 24 February 2006

Abstract

Background. Patient-based similarity metrics are important case-based reasoning tools which may assist with research and patient care applications. Ontology and information content principles may be potentially helpful tools for similarity metric development.

Methods. Patient cases from 1989 through 2003 from the Columbia University Medical Center data repository were converted to SNOMED CT concepts. Five metrics were implemented: (1) percent disagreement with data as an unstructured "bag of findings," (2) average links between concepts, (3) links weighted by information content with descendants, (4) links weighted by information content with term prevalence, and (5) path distance using descendants weighted by information content with descendants. Three physicians served as gold standard for 30 cases.

Results. Expert inter-rater reliability was 0.91, with rank correlations between 0.61 and 0.81, representing upper-bound performance. Expert performance compared to metrics resulted in correlations of 0.27, 0.29, 0.30, 0.30, and 0.30, respectively. Using SNOMED axis Clinical Findings alone increased correlation to 0.37.

Conclusion. Ontology principles and information content provide useful information for similarity metrics but currently fall short of expert performance.

Keywords: Datamining; Natural language processing; Similarity metrics; Electronic medical records; Ontology; Information content

1. Introduction

Improvements in data interchange standards, information systems, and data entry technologies have produced increased patient data in electronic format for clinical information systems. With more automation and creation of electronic health records, the amount of available coded patient data continues to expand. While most clinical databases are primarily used and designed for direct patient care, other important applications for these databases often lag behind, including clinical research and quality functions. Medical informatics data mining research can potentially help to improve the use of clinical databases for these applications. With respect to data mining of patient data, most reported studies use small, manually edited databases in order to examine focused and specific clinical questions [1–3]. Many believe, however, that one of the great potentials of health care data mining is the automated utilization of real-time large clinical databases for these applications [4,5].

Similarity is a fundamental concept that can help with the task of automated information integration. Measures of similarity aim to assess the degree of closeness between

☆ This research was conducted at the Columbia University Department of Biomedical Informatics.

[*] Corresponding author. Fax: +1 202 772 2429.
E-mail address: gmelton@jhmi.edu (G.B. Melton).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.01.004

cases of interest and produce medically relevant interpretations of patients so that analogous case(s) can be discovered. Useful metrics aim to give a quantifiable evaluation as to how similar two patients are to one another, based upon the context of the question at hand.

Similarity can also be examined from different perspectives. Depending upon the application, a clinical researcher may be interested in identifying sets of similar patients, such as patients who might qualify for an experimental oncology chemotherapy protocol. Furthermore, in theory, similarity metrics may be able to generate new knowledge about how patients relate to one another. Some of the obvious categories of factors that might be important according to the question being answered include: a patient's disease or condition (diagnoses), patient intervention (procedure), and patient environment and pre-existing state (demographics, address).

In clinical medicine, similarity metrics are typically based upon the case, defined as one or several medical encounters, or the entire patient record. If metrics can approach the behavior of experts, they will likely have an important role in case-based reasoning for research and for real-production systems, including decision-support applications. We were interested in exploring the methodological issues with inter-patient distance metrics by examining case-based similarity from a general clinical perspective and in using ontology principles as a tool to examine previously described distance metrics in medical informatics and to construct additional measures. These measures were then evaluated on patient cases with expert evaluation as our gold standard.

2. Background

2.1. Similarity metrics in clinical medicine

Similarity metrics in the general literature have most commonly focused on *semantic similarity*. Semantic similarity metrics quantify similarity in meaning between two concepts. Reported measures have used techniques of mutual information [6], Dice coefficient [7], cosine coefficient [7], information content [8], and distance-based metrics [9,10].

Semantic similarity has also been extended beyond single concept level comparisons to comparisons of two words with multiple meanings. For this extended case, Resnik [8] proposed that the task of semantic similarity between words could be dealt with in one of two ways: (1) by comparing all term pair combinations and weighting the more frequent senses more heavily or (2) by taking the maximum similarity between each term sense for one word and the closest term sense for the other word. In an analogous problem to the second approach, metrics between cases (*inter-patient similarity*) also compare large *sets* of data composing each case, where each piece of data in one case is measured relative to the closest data element in the other patient in order to derive an overall metric.

The study of similarity can also be viewed in terms of the information sources used. Knowledge-free approaches can be implemented, which rely on statistical measures, such as term frequency and co-occurrence data [11]. As medicine is a large and complex domain which is rich in synonymy and rich in semantically similar or related concepts, these knowledge-free approaches are often not ideal. Taxonomies and ontologies provide an information-rich framework from which to compare concepts within a domain. An ontology has a structured format with relationships between concepts. The "isa" relationship with parent and child is the core relationship, and other semantic relationships provide additional associations between terms (such as "part-of" or "active-ingredient-of"). While an important strength of ontologies is their information-rich nature, one of the major drawbacks arises from the inconsistent completeness in many terminologies with respect to structure and content. In particular, with respect to medical ontologies, concept coverage for clinical sub-domains can vary markedly for different terminologies [12].

With semantic similarity and ontologies, Caviedes and Cimino [13] proposed a distance metric between biomedical concepts using the minimum number of "isa" parent links between two target concepts in the Unified Medical Language System (UMLS). Lord et al. [14] applied a similarity score using Gene Ontology (GO) and information content to measure similarity between GO entries. While these studies focus on semantic distance as opposed to inter-patient distance, they demonstrate a foundation to use ontology principles to assess inter-patient distance. To date, most reports in the clinical literature with patient-based distance have used structure-free approaches or have focused upon a medical subspecialty with relatively small numbers of data variables [15,16].

2.2. SNOMED CT terminology

The information-rich framework provided by medical terminologies with the relationships provided between concepts led us to look closely at several clinical terminologies as tools for case-based distance metrics. Some of the candidate terminologies considered included ICD9-CM, SNOMED CT, and UMLS. As medical terminologies have structured formats with relationships between concepts, we were interested in a terminology with good concept coverage and with a rich structure of relationships between concepts, including "isa" relationships and other semantic relationships.

ICD9-CM, commonly used for administrative coding, deals exclusively with diagnosis and procedure related concepts. Relationships between concepts are inferred and not always complete or accurate in meaning. Each diagnosis is classified into disease-specific categories, and each of these has a subcategory diagnosis from which a specific diagnosis code is derived. While ICD9-CM does give important relationships of organ systems and some classes of diagnoses, it does not have a more extensive "isa" type hierarchy from its structure. In addition, the ICD9-CM structure contains significant inaccuracies and inconsistencies [17,18]. ICD9-CM also does not contain concepts unrelated to diagnosis or procedure. Furthermore, there are important issues associated with the consistency, quality, and accuracy of administrative coding [19–22].

Of the various terminologies in the medical domain, the UMLS contains the greatest number of medical concepts, as it is a conglomeration of many different terminologies. While the UMLS preserves semantic relations as expressed in the source terminologies, it does not intrinsically have a well-structured set of relationships between concepts which would be desired for our study of similarity. Thus, the UMLS itself does not have the desired features from a terminology which we sought for applying our similarity metric.

SNOMED CT was the terminology implemented in this study for several important reasons. First, the terminology is now publicly available. Second, the terminology has proven to have very good concept coverage [12,23–26] with over 361,800 concepts as of July 2004. Third, the defining relationships of SNOMED CT include an extensive "isa" structure along with containing other semantic relationships. From the top of the SNOMED CT hierarchy, 18 separate classes of concepts have been defined in SNOMED CT. SNOMED CT also has 46 different semantic relationship pairs in addition to the "isa" and "inverse-isa" relationship pair, which may be used to relate two concepts to one another. Conceptually, the SNOMED CT terminology may be considered a single tree with the concept "SNOMED CT concept" at the root with 18 children representing the 18 classes or axes of concepts. Alternatively, it may be thought of as 18 separate trees with semantic links between trees. The 18 SNOMED CT axes were also examined as possible similarity features in our evaluation.

2.3. Inter-patient distance considerations

Semantic distance measures the relative closeness between two concepts of interest from a terminology or concept-oriented view. *Inter-patient distance* compares the relative closeness between two cases (sets of patient data) of interest. *Clinical distance*, used in the calculation of inter-patient distance, is the amount of relative evidence for closeness from an inter-patient distance perspective when comparing a single concept in one case with the nearest concept in a second case. The word "clinical" is used for this quantity because the magnitude of the distance between concepts from a case-based view is influenced by the clinical granularity of the concept(s) in question (see Consideration 2). There were several other important observations which were noted.

Consideration 1. Inter-patient distance is *not* determining if cases are identical.

Rather, inter-patient distance attempts to quantify: given the information known about two cases, how much *evidence* is there that two patients are similar? Two patients can have the same or a very similar description and not be the same patient, particularly when the amount of data describing the patient is minimal. Patients can therefore have a distance metric to themselves greater than zero if there is minimal evidence for similarity.

Consideration 2. Semantic distance between two concepts is *different* than clinical distance between two case features.

Whether the difference between semantic and clinical distance is meaningful in computing an overall inter-patient distance metric remains unclear, as patient cases are typically composed of many coded data assertions, and differences at the individual term level may result in having minimal impact on the overall metric. As previously stated, we use the concept of "clinical" for this quantity because the magnitude of the distance between concepts from a case-based view is influenced by the clinical granularity of the concept(s) in question, as the level of abstraction or generalization of terms is important.

Using the example in Table 1 with concepts "heart disease" (HD) and "mitral valve prolapse" (MVP), these ideas concerning concept generalization can be illustrated with semantic versus clinical distance. The concepts in examples 1 and 2 have a semantic distance of zero between one another: $\text{Dist}_{\text{sem}}(\text{HD},\text{HD}) = 0$ and $\text{Dist}_{\text{sem}}(\text{MVP},\text{MVP}) = 0$, which makes both pairs equally close to one another. From a case-based perspective, however, patient A and B in example 2 with MVP have more evidence of similarity than the patients with HD in example 1: $\text{Dist}_{\text{Clin}}(\text{HD},\text{HD}) > \text{Dist}_{\text{Clin}}(\text{MVP},\text{MVP})$. This is because while both patients with HD from a semantic perspective have the same concept and therefore a semantic distance of zero, when applying these concepts to the patient case, HD could mean many other disease entities, including MVP, coronary artery disease, congestive heart failure, aortic stenosis, and others whereas MVP refers to the same particular disease entity. We would also assert that in example 3, patient A with HD and patient B with MVP (example 3) have a distance of a comparable magnitude to the two patients in example 1 that both have HD: $\text{Dist}_{\text{Clin}}(\text{HD},\text{HD}) \approx \text{Dist}_{\text{Clin}}(\text{HD},\text{MVP})$. While it is possible that both patients have MVP, since HD subsumes many additional disease entities, statistically it is likely that these two patients have two different varieties of HD.

The clinical distance should therefore be influenced by the closeness between concepts within a terminology and also by the informativeness of the nodes at issue. As such, leaf nodes are more informative and specific than non-leaf nodes, which can provide more potential evidence or lack

Table 1
Sets of example patients with mitral valve prolapse (MVP) and heart disease (HD)

	Example 1	Example 2	Example 3
Patient A	HD	MVP	HD
Patient B	HD	MVP	MVP

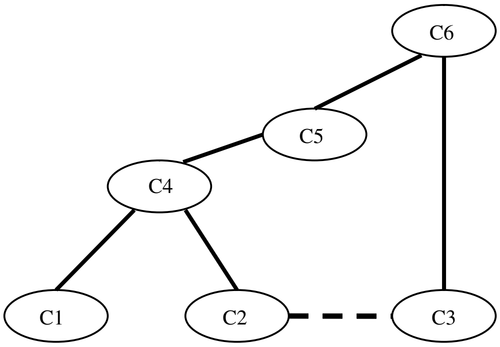

Fig. 1. Ontology with "isa" links (solid line) and non-"isa" links (broken line). C4 subsumes C1 and C2. The concept that subsumes C2 and C3 is less clear if non-"isa"links are used.

of evidence for similarity between cases. As a result, nodes with more evidence for similarity should contribute more to the overall distance metric, which is typically taken into account by using a weighting factor.

One method described previously is the use of information content from information theory principles [8], which for a concept is defined as $I(\text{concept}) = -\log P(\text{concept})$. Here, $P(\text{concept})$ is the probability of concept, which can be calculated using a variety of methods, including corpus statistics. In an "isa" taxonomy, the concept that subsumes the two concepts of interest is used to measure the similarity between concepts. In Fig. 1, if C1 and C2 were compared, the information content of C4 would be a metric reflecting the information shared by the concepts within the "isa" hierarchy. In a hierarchy with non-"isa" links (depicted as dashed line in Fig. 1), if C2 and C3 were being compared, C6 would be the correct concept to use if "isa" links were only considered. If non-"isa" links are used, then the correct concept(s) to use becomes less clear.

Consideration 3. Clinical distance from a concept in one case to the nearest concept in another case can be calculated using defined relationships to find the "minimal-cost" path.

From Consideration 2, it follows that a non-zero cost would be associated with the starting node itself. The number of path links between concepts does not achieve this property, but it is a useful comparison. The number of links between nodes is also highly dependent upon the pre-defined network hierarchy. Two approaches that might be less sensitive to this issue include the use of a prevalence measure for cost and/or the use of information content. Nodes that are less granular (many descendants) should have a higher cost than those more granular (few descendants).

3. Methods

3.1. Data preparation

Patient demographic data, coded procedure and diagnosis data (ICD9-CM codes), and text documents from the Columbia University Medical Center (CUMC) clinical data repository from 1989 through 2003 were used for this study (Fig. 2). Text documents included discharge summaries, clinic notes, transfer of service notes between physician housestaff, radiology, pathology, microbiology documents, and other ancillary reports. Documents not available in electronic format were daily progress notes, nursing notes, and some ambulatory care notes. Although medication and laboratory data were sometimes included in summary form in text documents, such as discharge summaries and admission notes, pharmacy and laboratory data were excluded from the dataset.

Data were converted to SNOMED CT codes using automated tools. Demographic data were converted with a script converting data to SNOMED CT codes. ICD9-CM codes were converted to SNOMED CT concepts using tools from the UMLS (Version 2004AB, file MRCONSO). Text documents were encoded to structured target terms using the natural language processor MedLEE (Medical Language Extraction and Encoding System) [27]. As previously reported, MedLEE has a module which maps its target terms in an automated manner to UMLS codes [28]. Negative or uncertain MedLEE assertions were then eliminated and UMLS tools (file MRCONSO.RRF) were used to convert each UMLS code to SNOMED CT. Each patient case was ultimately represented as a set of positive SNOMED CT coded assertions.

3.2. Distance metrics

We decided to apply a series of distance metrics to evaluate inter-patient distance. For the first metric, data was viewed as an unstructured "bag of findings," and the proportion of disagreement between cases was calculated (Eq. (1)).

$$\text{Dist}_{\text{Inter-pt}} = \frac{\text{Number of features in one but not both cases}}{\text{Number of features in either case}}. \tag{1}$$

From a mathematical perspective, this metric is a true distance in that the distance from a patient to himself is zero and that the metric is both communicative and associative. Metric 1 is a knowledge-free approach.

The remaining four distance metrics used the SNOMED CT defining relationships. When calculating shortest path between two nodes for these metrics, the number of non-"isa" or non-"inverse-isa" links was limited to one. In this manner, concepts with important relationships between two patients might be discovered, such as a particular disease and its organ system. In addition, limiting the number of non-"isa" or non-"inverse-isa" links to one was helpful in limiting unrelated or non-sense relationship paths from being found. No limit on the number of "isa" or "inverse-isa" links was placed so that the least upper bound distance between any two concepts would potentially be traversing directly from the first concept to the top node ("SNOMED CT concept") and down to the second node (Fig. 1).

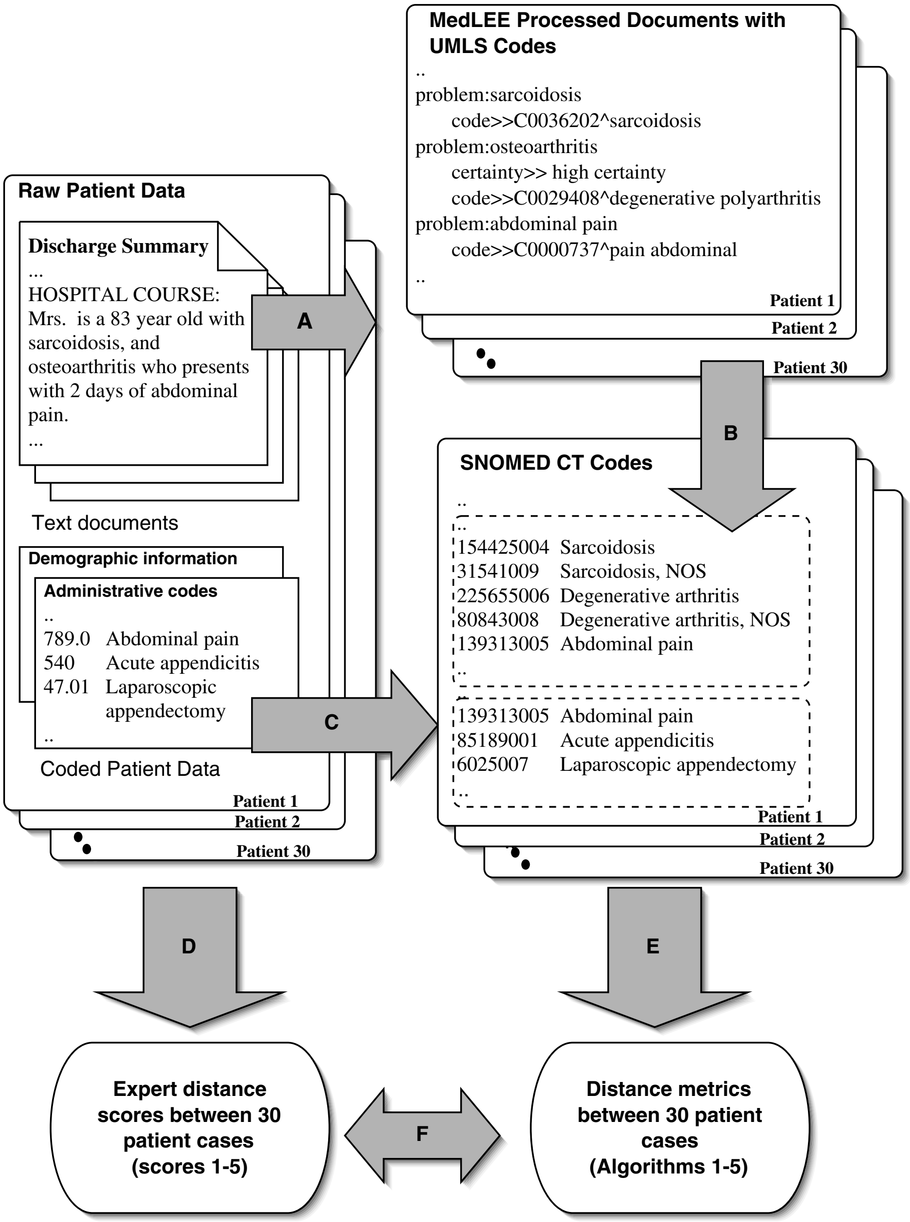
MedLEE Processed Documents with UMLS Codes
..
problem:sarcoidosis
code>>C0036202^sarcoidosis
problem:osteoarthritis
certainty>> high certainty
code>>C0029408^degenerative polyarthritis
problem:abdominal pain
code>>C0000737^pain abdominal
..
Patient 1
Patient 2
Patient 30
Raw Patient Data
Discharge Summary
...
HOSPITAL COURSE:
Mrs. is a 83 year old with sarcoidosis, and osteoarthritis who presents with 2 days of abdominal pain.
...
A
Text documents
Demographic information
Administrative codes
..
789.0 Abdominal pain
540 Acute appendicitis
47.01 Laparoscopic appendectomy
..
Coded Patient Data
Patient 1
Patient 2
Patient 30
B
SNOMED CT Codes
..
..
154425004 Sarcoidosis
31541009 Sarcoidosis, NOS
225655006 Degenerative arthritis
80843008 Degenerative arthritis, NOS
139313005 Abdominal pain
..
139313005 Abdominal pain
85189001 Acute appendicitis
6025007 Laparoscopic appendectomy
..
C
Patient 1
Patient 2
Patient 30
D
E
Expert distance scores between 30 patient cases (scores 1-5)
F
Distance metrics between 30 patient cases (Algorithms 1-5)

Fig. 2. Experimental design flow diagram of inter-patient distance metric evaluation. Example patient data snippets included for illustrative purposes. Arrow A, text document data is converted to coded format with UMLS codes using MedLEE. Arrow B, UMLS codes are converted to SNOMED CT with UMLS tools (MRCONSO). Arrow C, coded data from raw repository converted to SNOMED CT codes using automated and manual tools. Arrow D, raw repository data evaluated by experts using WebCIS clinical information system to view data to derive distance scores between cases. Arrow E, Algorithms 1–5 used on SNOMED CT codes from patient cases to derived distance metrics distance metrics between the 30 cases. Arrow F, expert scores and distance metrics compared statistically.

The overall distance metric between two patients is then the average D from the first to the second patient and D from the second to first patient (Eq. (2)). From this, the overall metric is communicative. As detailed in Consideration 1 the inter-patient distance from a patient to himself would be greater than zero. These metrics are also not generally associative.

$$\mathrm{Dist}_{\text{Inter-pt}}(P_{\mathrm{A}}, P_{\mathrm{B}}) = \mathrm{Dist}_{\text{Inter-pt}}(P_{\mathrm{A}}, P_{\mathrm{B}}) = \frac{D_{P_{\mathrm{A}},P_{\mathrm{B}}} + D_{P_{\mathrm{B}},P_{\mathrm{A}}}}{2}. \tag{2}$$

In general, $D_{P_{\mathrm{A}},P_{\mathrm{B}}} \neq D_{P_{\mathrm{B}},P_{\mathrm{A}}}$, where P_{A} is patient A and P_{B} is patient B. $D_{P_{\mathrm{A}},P_{\mathrm{B}}}$ is the weighted average of clinical distance between each of the s nodes (v_i) in patient A to the node of patient B (v_b) which minimizes the clinical distance, and where w_i is the weight of the ith node of patient A (Eq. (3)). Because the quantity $D_{P_{\mathrm{A}},P_{\mathrm{B}}}$ is reliant upon each of P_{A}'s nodes in relation to the closest node of P_{B}, this quantity may be very different than $D_{P_{\mathrm{A}},P_{\mathrm{B}}}$.

$$D_{P_{\mathrm{A}},P_{\mathrm{B}}} = \frac{\sum_{i=1}^{s} w_i \min_{v_b \in P_{\mathrm{B}}} [\mathrm{dist}_{\mathrm{clin}}(v_i, v_b)]}{\sum_{j=1}^{s} w_j}. \tag{3}$$

The second metric was average minimal number of links between concepts. Using Eq. (3), $w_i = 1$ and $\mathrm{dist}_{\mathrm{clin}}(v_i, v_b) =$ number of path links. This metric implements the shortest path problem and is similar to other previously described measures that use path length [9,10,13]. As alluded to previously, this metric does not take into account Consideration 2, since the number of links from a concept to itself is zero.

The third metric expands upon the second metric by using information content principles and keeps the clinical distance as the number of path links. The number of descendants of the term (number of descendants of the term, including the term itself) was used to construct a weighting factor. While Resnik and others consider only "isa" links and use the superconcept that subsumes the two concepts at issue, we considered other semantic relationships in addition to "isa" links. As a result, the probability of the superconcept within the "isa" hierarchy would be artificially high when another semantic relationship is considered, and the weighting factor would become very small. For this reason, we used the term at issue alone for our weight. Another issue with using the number of descendants is that this quantity is dependent upon SNOMED CT hierarchy variations.

$$w_i = -\log P(v_i),$$
$$\text{where } P(v_i) = \frac{\text{Number of idesc}(v_i)}{\text{Total number of terms}} \quad \text{and}$$
$$\mathrm{idesc}(v_i) = \text{descendents of } v_i, \text{including } v_i. \tag{4}$$

The fourth metric modifies the weighting factor by using term frequency to calculate information content. By using term frequency instead of descendents, there are theoretically less issues due to SNOMED CT hierarchy variations. The clinical distance remains as the number of path links. Term frequency for a concept was based on the frequency of the concept or any of its descendant concepts in the SNOMED CT "isa" hierarchy per patient. Term frequency was calculated on each of all SNOMED CT concepts on a random set of 14,204 (N) patients in the clinical data repository.

$$w_i = -\log P(v_i), \quad \text{where } P(v_i) = \frac{\sum_{n \in \mathrm{idesc}(v_i)} \mathrm{count}(n)}{N}. \tag{5}$$

The fifth and final metric combines information content with a minimal cost path between concepts with variable node weights. Because the distance of a node to itself is not zero with this measure, this metric is better referred to as a "dissimilarity metric" instead of a "distance," as it uses Considerations 2 and 3. Using the same weighting factor with descendants (metric 3, Eq. (4)), the clinical distance between nodes was calculated by summing the proportion of descendants for the x nodes including the starting node, the ending node, and any node associated with a change in direction along the path or change in link-type (Eq. (5)).

$$\mathrm{dist}_{\mathrm{clin}}(v_i, v_b) = \sum_{y=1}^{x} P(v_y),$$
$$\text{where } P(v_y) = \frac{\text{Number idesc}(v_y)}{\text{Total number of terms}}. \tag{6}$$

Contrasting the four structured metrics to one another, the second metric is the most straightforward. It uses average minimal number of links between concepts and the defined relationships to calculate semantic distances between terms. The third metric is similar to metric two, as path distance is the same, but it expands upon the second metric by using information content principles for the weight. This metric utilizes the defined relationships and term depth (number of idescendants) to take into account the specificity of each term along the path. The fourth metric also uses information content principles for the weight, but instead uses term frequency within the corpus in question instead of descendents. This is metric is theoretically less influenced by SNOMED CT hierarchy variations. The fifth and final metric, in comparison to all of the other metrics, combines information content with a clinical distance. By using a clinical distance, this metric takes into account all of the discussed considerations, as it uses information content for the weight, as well as implements a clinical distance to construct the distance between nodes of one patient to nodes of a second patient.

3.3. Metric evaluation

The five distance metrics were evaluated using 30 random patient cases with an inpatient hospitalization admitted to any of the six general medical-surgical inpatient floors at CUMC in 2003 (Fig. 2). Metrics were computed overall and along each of the 18 SNOMED CT axes (e.g., Procedure, Specimen, Events). Much like the experiment often conducted for word similarity [29], our gold standard was expert evaluation by three physicians (G.B.M., F.P.M., A.S.R.). For this study, the definition of similarity was vague, and evaluators were asked to rate case similarity based upon what made clinical sense to them. Experts used the electronic medical record browser WebCIS at CUMC [30] to access the entire electronic patient chart from 1989 through 2003, which was the source data from which the encodings for each of the algorithms were derived. The electronic chart consisted of discharge summaries, operative notes, radiology reports, pathology reports, ICD9-CM diagnosis coding, laboratory values, and other procedure reports. Experts assigned case pairs a score between 1 and 5, with 1 indicating "identical/very high degree of similarity" and 5 meaning "not at all similar." Inter-rater reliability was calculated using Cronbach's Alpha [31]. Expert evaluations were correlated to one another, as were the inter-patient distance measures to average expert scores, using the Spearman's rho coefficient of rank correlation [32]. To estimate the variance of the correlation coefficient and their differences, a bootstrap algorithm was implemented [33], and then a normal approximation was used to estimate p values.

4. Results

The three physician experts rated 30 cases with an overall inter-rater reliability 0.91. The correlation amongst expert rater scores ranged from 0.61 to 0.81, which represents an upper bound for automated performance. There was no statistical difference in the correlation amongst the experts. While the entire electronic chart was available to the experts, for the task of rating case similarity based upon what made clinical sense, all three experts reported that discharge summaries were predominantly used in their similarity score determinations. The second most common source used by the three experts was reported as ICD9-CM codes by one and operative notes by two of the experts.

The average expert rating for each pairwise comparison of the 30 cases was correlated with the results for each metric, as depicted in Table 2. The structured algorithms had a slightly greater correlation than when data were viewed as an unstructured "bag of findings" (Algorithm 1 (0.27) versus Algorithm 2 (0.29) or versus Algorithms 3, 4, or 5 (0.30)), but the differences were not significant. These correlations represent 30% higher of a likelihood that the metrics would agree than disagree with the experts. Furthermore, as expected, experts were statistically more like to agree with one another than any of the algorithms were likely to agree with the experts ($p < 0.0001$).

Table 3
Correlation of data from individual SNOMED CT axes and average expert scores with Algorithms 2 and 5

Axis	Metric (2) minimum number of links	Metric (5) descendant weight, descendant path cost
Clinical Finding	0.34	0.37
Body Structure	0.29*	0.28***
Procedure	0.22**	0.25****

* $p = 0.11$, "Clinical Finding" metric 2 versus "Body Structure" metric 2.
** $p = 0.0015$, "Clinical Finding" metric 2 versus "Procedure" metric 2.
*** $p = 0.035$, "Clinical Finding" metric 5 versus "Body Structure" metric 5.
**** $p = 0.036$, "Clinical Finding" metric 5 versus "Procedure" metric 5.

Each term was also categorized into its SNOMED CT axis to correlate the axes to expert scores. The metrics were re-calculated along each of the 18 SNOMED CT axes with the second and fifth algorithm. As depicted in Table 3, the axes that correlated most strongly with experts were axes "Clinical Finding," "Body Structure," and "Procedure" for algorithms 2 and 5, respectively. "Clinical Finding" for algorithm 2 was significantly more likely to agree with experts than the basic algorithm 2 ($p = 0.0080$) or "Procedure" for algorithm 2 ($p = 0.0015$) but not more likely to agree than "Body Structure" for algorithm 2 ($p = 0.11$). Similarly, "Clinical Finding" for algorithm 5 was significantly more likely to agree with experts than the basic algorithm 5 ($p = 0.0025$), "Procedure" for algorithm 5 ($p = 0.036$), and "Body Structure" for algorithm 5 ($p = 0.035$).

Table 2
Metric performance correlation with average expert ratings

Metric	Detailed metric description[a]	Correlation to experts
Experts		0.61–0.81[b]
(1) Bag of findings	$\text{Dist}_{\text{Inter-pt}} = \frac{\text{Number of features in one but not both cases}}{\text{Number of features in either case}}$	0.27
(2) Average links between concepts	$w_i = 1$ and $\text{dist}_{\text{clin}}(v_i, v_b) =$ number of path links	0.29
(3) Links weighted by information content with descendants	$w_i = -\log P(v_i)$, where $P(v_i) = \frac{\text{Number of idesc}(v_i)}{\text{Total number of terms}}$ where $\text{idesc}(v_i) =$ descendents of v_i, including v_i and $\text{dist}_{\text{clin}}(v_i, v_b) =$ number of path links	0.30
(4) Links weighted by information content with term frequency	$w_i = -\log P(v_i)$, where $P(v_i) = \frac{\sum_{n \in \text{idesc}(v_i)} \text{count}(n)}{N}$ where $N =$ total number of patients and $\text{dist}_{\text{clin}}(v_i, v_b) =$ number of path links	0.30
(5) Path distance with descendants weighted by information content with descendants	$w_i = -\log P(v_i)$, where $P(v_i) = \frac{\text{Number of idesc}(v_i)}{\text{Total number of terms}}$ where $\text{idesc}(v_i) =$ descendents of v_i, including v_i and $\text{dist}_{\text{clin}}(v_i, v_b) = \sum_{y=1}^{x} P(v_y)$, where $P(v_y) = \frac{\text{Number idesc}(v_y)}{\text{Total number of}}$	0.30

[a] For algorithms 2 through 5,

$$\text{Dist}_{\text{Inter-pt}}(P_A, P_B) = \text{Dist}_{\text{Inter-pt}}(P_A, P_B) = \frac{D_{P_A,P_B} + D_{P_B,P_A}}{2}$$

and

$$D_{P_A,P_B} = \frac{\sum_{i=1}^{s} w_i \min_{v_b \in P_B} [\text{dist}_{\text{clin}}(v_i, v_b)]}{\sum_{j=1}^{s} w_j},$$

where D_{P_A,P_B} is weighted average of clinical distance between each of the s nodes (v_i) in patient A to the node of patient B (v_b) which minimizes the clinical distance, and where w_i is the weight of the ith node of patient A.

[b] Correlation of experts to one another.

5. Discussion

Inter-patient similarity metrics can potentially help datamining researchers answer complex clinical questions for large populations. The issues with this task remain challenging. This study represents a formal attempt to address and discuss some of the underlying issues with inter-patient distance using ontology and information content principles as tools. While the addition of ontology principles and information content did result in improved metrics, metric performance is still not as good as experts.

Despite the increase in observed correlation when comparing an unstructured "bag of findings" to the other four metrics, there was virtually no difference in the correlations in the four different metrics which used SNOMED CT. This was unexpected, as we anticipated that as algorithms progressed from less to more complex (algorithm 2 up to 5), that metric performance would improve and change as more of the described patient similarity principles were incorporated into the metrics. As such, while all the discussed inter-patient similarity principles appear to be important from an intuitive standpoint, which one(s) will be key for inter-patient similarity in practice has yet to be determined. While using descendants within the metrics would intuitively be influenced by intricacies and variations in SNOMED CT, using descendants versus term frequency did not have a negative effect upon metric performance in our experiment.

It is unclear what relative contribution inaccurate coding of data during conversion to SNOMED CT or the somewhat fuzzy structure of SNOMED CT itself had upon the metrics and our results. There was almost certainly some loss of granularity and accuracy in our results due to issues associated with converting from a source terminology (such as ICD9 codings) to SNOMED CT. Our results could potentially be improved if all data were initially encoded in SNOMED CT without the need to convert data from one source terminology to another. Clearly, the usefulness of ontology principles as tools for a particular purpose depends highly upon the quality and granularity of the terminology in question. As such, we speculate that some of the stakeholders in the development of SNOMED CT (pathology or internal medicine) may have improved the quality of concept and hierarchy development in these areas more than those domains with less traditional involvement with SNOMED CT's development (public health or surgery).

We were able to demonstrate that patient similarity assessment is well correlated among expert physician raters given the task of defining similarity as what made clinical sense to them. Similarity for this set of random patient cases appears to correlate most strongly with concepts within the SNOMED CT axes "Clinical finding," "Body Structure," and "Procedure," which roughly correspond to (1) diseases, signs, and symptoms, (2) organ systems, and (3) procedures, respectively. This is perhaps somewhat intuitive but is an interesting observation to formally identify. An interesting follow-up study could be to formally conduct a cognitive analysis of experts with the task of case-based similarity from several perspectives. In this manner, we may be able to understand the cognitive process used by experts and how this changes according the question or task at hand in order to optimize future computer performance. Furthermore, while ICD9-CM was rejected for several of its limitations from a structural and content coverage perspective, it would be interesting to use ICD9-CM in a follow-up experiment to compare its performance to SNOMED CT, as ICD9-CM's coverage of these concept class axes is good. As a follow-up experiment, all five of these metrics could be applied to data from the discharge summaries alone, which was thought by experts to be the highest yield data source. This could help to test the degree to which incomplete SNOMED coding, or other information associations, contributed to the poorer observed performance of the metrics.

In summary, the authors propose several patient-based distance measures with varying complexity, building upon previously described measures. Our evaluation of these measures reveals that useful information can be obtained from these techniques, but further research will be needed in order to achieve the goal of automated utilization of these measures for practical uses. Future work should aim to understand better these correlations, to leverage this knowledge to create better metrics, and to identify subsets of patient data best correlated with experts.

6. Conclusion

Inter-patient similarity is a fundamental area of development in datamining research. Defined relationships of the terminology and principles of information content were able to provide valuable information for distance metrics, and narrowing in on the features used in expert determination of similarity with SNOMED CT axes was helpful. These measures, however, currently fall short of expert performance.

Acknowledgments

This work was funded by National Library of Medicine (NLM) "Discovering and applying knowledge in clinical databases" (R01 LM06910). Drs. Melton, Morrison, and Rothschild were supported by NLM Training Grant (5T15LM007079-12). We thank Carol Friedman for the use of MedLEE (NLM support R01 LM06274 and R01 LM07659).

References

[1] Coulter DM, Bate A, Meyboom RH, Lindquist M, Edwards IR. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. BMJ 2001;322(7296):1207–9.

[2] Harris Jr JM. Coronary angiography and its complications. The search for risk factors. Arch Intern Med 1984;144(2):337–41.

[3] Markey MK, Lo JY, Tourassi GD, Floyd Jr CE. Self-organizing map for cluster analysis of a breast cancer database. Artif Intell Med 2003;27(2):113–27.

[4] Bellazzi R, Zupan B. Intelligent data analysis–special issue. Methods Inf Med 2001;40(5):362–4.

[5] Safran C. Using routinely collected data for clinical research. Stat Med 1991;10(4):559–64.

[6] Hindle D. Noun classification from predicate-argument structures. In: Proceedings of the ACL-90, 1990, Pittsburgh, Pennsylvania; 1990. p. 268–75.

[7] Frakes WB, Baeza-Yates R, editors. Information retrieval, data structure and algorithms. New Jersey: Prentice Hall; 1992.

[8] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on AI; 1995. p. 448–53.

[9] Lee JH, Kim MH, Lee YJ. Information retrieval based on conceptual distance in is-a hierarchies. J Doc 1989;49(2):188–207.

[10] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;19(1):19–30.

[11] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference on research in computational linguistics, 1997; 1997. p. 19–33.

[12] Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes and Structures. J Am Med Inform Assoc 1996;3(3):224–33.

[13] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform 2004;37(2):77–85.

[14] Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003;19(10): 1275–83.

[15] Hamilton PW, Bartels PH, Anderson N, Thompson D, Montironi R, Sloan JM. Case-based prediction of survival in colorectal cancer patients. Anal Quant Cytol Histol 1999;21(4):283–91.

[16] Tsatsoulis C, Amthauer HA. Finding clusters of similar events within clinical incident reports: a novel methodology combining case based reasoning and information retrieval. Qual Saf Health Care 2003;12(Suppl. 2):ii24–ii32.

[17] Cimino JJ. An approach to coping with the annual changes in ICD9-CM. Methods Inf Med 1996;35(3):220.

[18] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4–5):394–403.

[19] Malenka DJ, McLerran D, Roos N, Fisher ES, Wennberg JE. Using administrative data to describe casemix: a comparison with the medical record. J Clin Epidemiol 1994;47(9):1027–32.

[20] Preen DB, Holman CD, Lawrence DM, Baynham NJ, Semmens JB. Hospital chart review provided more accurate comorbidity information than data from a general practitioner survey or an administrative database. J Clin Epidemiol 2004;57(12):1295–304.

[21] Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med 1988;318(6):352–5.

[22] Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Med Care 2004;42(11):1066–72.

[23] Lieberman MI, Ricciardi TN, Masarie FE, Spackman KA. The use of SNOMED CT simplifies querying of a clinical data warehouse. AMIA Annu Symp Proc 2003:910.

[24] Penz JF, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, et al. Evaluation of SNOMED coverage of Veterans Health Administration terms. Medinfo 2004;11(Pt. 1):540–4.

[25] Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthc Inform 2004;21(9):54–6.

[26] Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc 2003:699–703.

[27] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.

[28] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402.

[29] Miller GA, Charles WG. Contextual correlates of semantic similarity. Lang Cognit Process 1991;6(1):1–28.

[30] Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large scale deployment of a Web-based clinical information system. Proc AMIA Symp 1999:804–8.

[31] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

[32] Gibbons JD. Nonparametric methods for quantitative analysis. 3rd ed. Ohio: American Sciences Press; 1997.

[33] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall; 1993.