An improved clinical data similarity algorithm based on ICD10

Wenjun Yu, Kai Wang, Bin Hu and Yu Huang*

College of informatics, Huazhong Agricultural University, Wuhan, Hubei, China

*Corresponding author: yhuang@mail.hzau.edu.cn

Abstract. Disease prediction helps prevent disease and early diagnosis, and accurate classification of patients greatly improves the accuracy of disease prediction. Today's massive multi-dimensional medical data and its similarity algorithms provide the basis for the classification of clinical diseases. Based on this, we randomly generated the simulated clinical data of ICD10 structure, used the improved similarity algorithm to calculate the similarity and classification of the two patients, and found the patients belonging to different disease categories in the classified patient group. This finding provides a scientific basis for the correction of genetic algorithms and genetic research.

1. Introduction
Precisely predicting the patient's disease development status can help prevent disease, early diagnosis and timely intervention. It can not only control the growth of high medical costs, but also avoid or slow down many people's disease, so it has important social and scientific significance. [1]. In the past, doctors mainly judged the patient's future development based on experience, but with the substantial increase of clinical data, it is more and more difficult for clinicians to rely solely on the human brain to predict the patient's condition. The accumulation of medical big data provides a new basis for predicting the patient's condition through medical data mining, which will promote medical care from the development of skilled art to data-driven science, from "thousands of people” to “personalized treatment", from the passive mode of "governing the disease” to the active mode of "treating the disease.”

At present, there are many kinds of prediction methods based on medical big data, which can be divided into two categories: One is to generate a prediction model based on data [2-4], that is, a supervised machine learning method for specific purposes, such as classification and deep neural networks, to generate training models to predict clinical events in unknown patients; the second is to measure the distance between patients, establish a similar group of patients, and predict the target patients through the characteristics of similar groups, which is called patient similarity analysis [5-7]. Specifically, patient similarity analysis refers to the selection of clinical concepts (such as diagnosis, symptoms, examinations, drugs, surgery, genes.) as characteristic items of patients in a specific medical environment, and quantitative analysis of complex conceptual semantic spaces. The distance between concepts, thereby dynamically measuring the distance between patients, and screening out similar groups of patients similar to index patients. In the medical big data scenario, the various features of the patient-like group can theoretically provide multiple predictions and have better universal characteristics than the training model for specific goals.

In general, patient similarity is not limited to predictive goals and the number of known patients with labels. The core method is “distance between concepts in complex conceptual semantic spaces” and
"measuring the distance between patients using the clinical concept of patients". It is unrelated and non-
supervised. It is a universal and measurable method. With the accumulation of big data resources and
the maturity of big data analysis technology, it will show an increasingly broad application scenario.

In this study, 12,000 patient data were simulated based on clinical data characteristics and encoded
in ICD10 format. An improved similarity algorithm was used to calculate the similarity of the two
patients and classified according to other similarity algorithms. Comparing the results, our classification
results are closer to the results of manual classification.

In this study, 12,000 patient data were simulated based on clinical data characteristics and encoded
in ICD10 format. An improved similarity algorithm was used to calculate the similarity of the two
patients and classification. Comparing with the results of other similarity classification algorithms, our
classification results were closer to those of manual classification.

2. Method

2.1 Data simulation and preprocessing
There are many types of clinical data. In general, all patient-related data can be used to analyze patient
similarity. The main source of these data is the data generated by clinical informationization, such as
test report, physical signs, medication, diagnosis, disease record and other information. The data is
divided into two categories: one is structured data stored in the database, such as numerical test results,
vital signs, demographic information, etc.; the other is unstructured data, such as disease records,
discharge records, check text information such as reports, as well as image and signal data such as CT,
X-ray, MRI, ultrasound and ECG.

Structured data has the inherent advantage of being suitable for computer processing and can meet
the computational needs after a specific transformation, so we code to generate structured clinical data,
map clinical concepts to coding, and construct new ones from given attributes. Attributes are added to
the attribute set, and the ICD10 code is assigned to the disease diagnosis result to facilitate subsequent
large-scale data analysis, as shown in Figure 1:

Figure 1. A total of 12,000 simulated clinical data encoded by ICD10.

2.2 Improved similarity algorithm
The clinical data we simulated is encoded by ICD10, which is a tree structure with 4 levels under the
virtual vertices, 22 first-level nodes (i.e. disease classification), and more than 12,000 four-level nodes
(i.e. disease names). When comparing the distances of two codes, the level-based method considers the
depth of the code and the shortest path between the two codes. The deeper the depth is, the stronger the
generalization of the code is, and the shorter the shortest path is, the closer the clinical significance of
the two is. Girardi compared probabilistic-based and tier-based distance metrics in the study, and the
results showed that the visualization of the former often brought patients with multiple diagnoses closer
to patients with a single diagnosis, and the visualization of the latter formed a clear distinction between
them. The boundary-based approach performs better [8].

After data preprocessing, the patient is expressed as a vector consisting of the values of the processed
clinical concept. Comparing the two patients is to compare the similarity degree of each of the same
data types and the same clinical significance in the vector one by one, and then combine them to analyze
the overall similarity of the two patients. Our idea is to measure the criticality of the patient's condition
with the sum of the squares of the features [9-12]. However, the two patients reaching the same criticality
may be through different “paths”, so the cosine of the feature vector is used to further correct the patient’s similarity. The generalized Mahalanobis distance between patient Xi and cluster C with means \( \mu = \{\mu_1, ..., \mu_n\} \) are defined as follows:

\[
d_p(X_i, C) = \sqrt{(X_i - \mu)^T S(X_i - \mu)}
\]

Equation1: We use \( X_i = [x_{i,1}, ..., x_{i,n}] \in \mathbb{R}^{i\times d} \) to represent the feature vector of patient i, where \( i = 1, ..., n \) and \( n \) is the number of patients and \( d \) is the number of features. \( y_i \) is the label assigned to the patient i with \( y_i \in \{1,2, ..., L\} \) and \( L \) is number of class labels and in our case, the number of medication plans. Medication plan is based on using drug or combination of drugs with specific dosages during the treatment.

Equation 2: Given two attribute vectors, \( A \) and \( B \), the remaining string similarity \( \theta \) is given by the dot product and the vector length, where \( A_i \) and \( B_i \) represent the components of vectors \( A \) and \( B \), respectively.

2.3 Classification
Patient similarity group analysis refers to the identification of patients who are similar to indexed patients. If a feature item that is of interest or related to an index patient (such as a complication € drug) is given a higher weight, the similar group can be further divided into a plurality of different sub-groups, called subsimilaromes [13]. With distance-based clustering methods, patients can be divided into several subsets, each of which is a cluster, referred to herein as a patient-like group. The clustering method can be divided into a partitioning method, a hierarchical method, etc.

Based on the simulated patient data characteristics, we use the hierarchical method of distance to create a hierarchical decomposition of a given patient set, which is formed into a cohesive or split form. In each successive iteration, one cluster is divided into smaller clusters until each patient is ultimately in a single cluster or a certain termination condition is met. The results show that the performance of k-means clustering and hierarchical clustering is quite good. Therefore, we use the TF-IDF method to assign weights to ICD10, use cosine values to calculate the distance between patients, and use the average connected hierarchical clustering method to divide patients into 189 sub-categories and explore the types and quantities of diseases included in each sub-category [14-17].

3. Result
Using the above result dataset, we analyzed how the choice of distance measurements affects the distance between 1549 cases when grouped into four primary (level 2) diagnoses. The distance matrices \( DJ \), \( DC \) and \( DH \) correspond to three choices of distance measurement \( dJ \), and \( dC \) and \( dH \) are respectively shown in Table 1.

### Table 1. Average distance between patients with grade IV (H) and appendix (A), gallbladder (G) and thyroid (T) disease. The three distance matrices use the Jaccard distance \( DJ \) (left), the Haase-Li conceptual distance \( DC \) (middle) and the new stratification distance \( DH \) (right).

|        | A     | H     | G     | T     |
|--------|-------|-------|-------|-------|
| \( DJ \) A H G T | 0.68  | 0.98  | 0.41  | 1.0   |
| \( DC \) A H G T | 0.25  | 0.83  | 0.21  | 0.87  |
| \( DH \) A H G T | 0.23  | 0.67  | 0.15  | 0.70  |

|        | A     | H     | G     | T     |
|--------|-------|-------|-------|-------|
|        | 0.99  | 0.98  | 0.98  | 1.0   |
|        | 0.85  | 0.41  | 0.17  |       |
|        | 0.70  | 0.25  |       |       |
|        | 0.98  | 0.52  |       |       |
4. Discussion
In this paper, we introduce a new similarity algorithm based on hierarchical distance measurement and compare it with two existing methods (Jaccard metric and Haase-Li conceptual distance) to calculate between two sets of concepts. The similarity in which a hierarchical arrangement can be used.

Our new hierarchical distance measurement allows the case to be grouped in ways that are not possible with the distance measurements available in the literature to date, and we extend the concept of established Jaccard distance by considering the concept hierarchy. If there is no conceptual hierarchy, our new distance measurement will revert to the Jaccard distance. Contrary to other level distance measurements, such as those proposed by Li et al. [18-20], our newly proposed method does not consider nonlinearity, so it can be said that the hierarchical information to the distance is more direct than other mappings.

Acknowledgments
This work was financially supported by the Fundamental Research Funds for the Central Universities (Program No. 2662015PY057) and the State Scholarship Fund (No. 201706765027).

References
[1] Belle A, Thiagarajan R, Soroushmehr S M, et al. Big data analytics in healthcare[J]. BioMed research international, 2015, 2015.
[2] Stiglic G, Brzan P P, Fijacko N, et al. Comprehensible predictive modeling using regularized logistic regression and comorbidity based features[J]. PLoS one, 2015, 10(12): e0144439.
[3] Wickramasinghe N. Deepr: a convolutional net for medical records[J]. 2017.
[4] Choi E, Bahadori M T, Sun J, et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism[C]//Advances in Neural Information Processing Systems. 2016: 3504-3512.
[5] Sharafoddini A, Dubin J A, Lee J. Patient similarity in prediction models based on health data: a scoping review[J]. JMIR medical informatics, 2017, 5(1): e7.
[6] Ng K, Sun J, Hu J, et al. Personalized predictive modeling and risk factor identification using patient similarity[J]. AMIA Summits on Translational Science Proceedings, 2015, 2015: 132.
[7] Brown S A. Patient similarity: emerging concepts in systems and precision medicine[J]. Frontiers in physiology, 2016, 7: 561.
[8] Girardi D, Wartner S, Halmerbauer G, et al. Using concept hierarchies to improve calculation of patient similarity[J]. Journal of biomedical informatics, 2016, 63: 66-73.
[9] Joshi R, Szolovits P. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2012, 2012: 1276.
[10] Panahiazar M, Taslimitehrani V, Pereira N L, et al. Using EHRs for heart failure therapy recommendation using multidimensional patient similarity analytics[J]. Studies in health technology and informatics, 2015, 210: 369.
[11] Köhler S, Schulz M H, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies[J]. The American Journal of Human Genetics, 2009, 85(4): 457-464.
[12] Wang B, Mezlini A M, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature methods, 2014, 11(3): 333.
[13] Jia Zheng, Zong Ruijie, Duan Huilong, et al. A Review on the Patient Similarity Analysis Based on Electronic Medical Records[J]. Chinese Journal of Biomedical Engineering, 2018, 37(3): 353-366.
[14] Roque F S, Jensen P B, Schmock H, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts[J]. PLoS computational biology, 2011, 7(8): e1002141.
[15] Verboon-Maciolek M A, Krediet T G, Gerards L J, et al. Severe neonatal parechovirus infection and similarity with enterovirus infection[J]. The Pediatric infectious disease journal, 2008, 27(3): 241-245.

[16] Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine[J]. Journal of biomedical informatics, 2011, 44(1): 118-125.

[17] Lee J, Maslove D M, Dubin J A. Personalized mortality prediction driven by electronic medical data and a patient similarity metric[J]. PloS one, 2015, 10(5): e0127428.

[18] Manjón J V, Coupé P, Buades A, et al. New methods for MRI denoising based on sparseness and self-similarity[J]. Medical image analysis, 2012, 16(1): 18-27.

[19] Sun J, Wang F, Hu J, et al. Supervised patient similarity measure of heterogeneous patient records[J]. ACM SIGKDD Explorations Newsletter, 2012, 14(1): 16-24.

[20] García M M, Allones J L I, Hernández D M, et al. Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations[J]. International journal of medical informatics, 2012, 81(8): 566-578.