International Journal of Medical Informatics 124 (2019) 97–103

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Measure clinical drug–drug similarity using Electronic Medical Records

Xian Zeng[b,1], Zheng Jia[b,1], Zhiqiang He[c], Weihong Chen[c], Xudong Lu[b], Huilong Duan[b], Haomin Li[a,d,*]

[a] The Children's Hospital, Zhejiang University School of Medicine, China
[b] College of Biomedical Engineering and Instrument Science, Zhejiang University, China
[c] Department of Pharmacy, Shanxi Dayi Hospital, China
[d] The Institute of Translational Medicine, Zhejiang University, China

ARTICLE INFO

Keywords:
Drug similarity
Electronic Medical Record
Drug clustering
Quantitative methods

ABSTRACT

Objective: Quantitative measurement of clinical drug–drug similarity has many potential applications in assessing medication therapy similarity and patient similarity. Currently, most of the methods to measure drug–drug similarity were not directly obtained from clinical data and cannot cover clinical drugs. We sought to propose a computational approach to measure clinical drug–drug similarity based on the Electronic Medical Record (EMR) system.

Materials and methods: We used the Bonferroni-corrected hypergeometric P value to generate statistically significant associations between drugs and diagnoses in an EMR dataset which contained 812 554 medication records and 339 269 discharge diagnosis codes. Then the Jaccard similarity coefficient was used to measure the distances between drugs. A k-means based bootstrapping method was proposed to generate drug clusters.

Results: The similarity matrix contains total 1210 clinical drugs used in the hospital was calculated. The clinical drug–drug similarity shows significant correlation with the chemical similarity of drugs and literature-based drug–drug similarity but with unique features. Based on this drug–drug similarity, 36 clinical drug clusters most of which were related to specific clinical conditions were generated. Detail of this drug clusters available at http://kb4md.org:4000/drugcluster.

Discussion: This method provided a whole new view of the relationship among clinical drugs. Furthermore, it has the potential to evaluate the effectiveness of drug knowledge translation and provide quantitative knowledge resources for many applications such as treatment comparisons and patient similarity.

Conclusion: We proposed a clinical drug–drug similarity measurement that generated from clinical practice data and covers all clinical drugs.

1. Introduction

Drug–drug similarity studies aim to find drugs which display similar pharmacological characteristics to the drug of interest and are driven by the hypothesis that similar drugs should be similar in mechanism of action, have similar side effect and be useful in treating a similar constellation of diseases [1]. The drug–drug similarity has extensive application in a variety of fields, such as drug repositioning [2,3], drug–drug interaction prediction [4,5], drug target identification [6] and drug side-effects prediction [7]. Except for these successful applications in pharmaceutical drug development, quantitative assessment of drug–drug similarity can also be used to analyze the clinical big data. Many clinical concepts such as diagnoses, symptoms, health behavior, signs and procedures have been applied to measure patient similarity and support the clinical decision. Quantitatively measure clinical drug–drug similarity can pave a way for medication therapy similarity and further patient similarity investigation.

Drug–drug similarity can be calculated from different resources. So far, several computational approaches which based on drug features such as chemical structure characteristics [8], gene expression profiles [9], side effect profiles [6,10], and biological target [2,11] have been successfully applied in drug–drug similarity analytics. On the other hand, medical literature is an additional source for measuring drug–-drug similarity. Brown et al. [1] extend methods to determine significantly co-occurring drug-MeSH term pairs in literature database and cluster drugs based on their pair-wise similarities. While all these

* Corresponding author at: No. 3333 Binsheng Road, Hangzhou, Zhejiang Province, 310052, China.
E-mail address: hmli@zju.edu.cn (H. Li).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.ijmedinf.2019.02.003
Received 23 May 2018; Received in revised form 21 December 2018; Accepted 10 February 2019

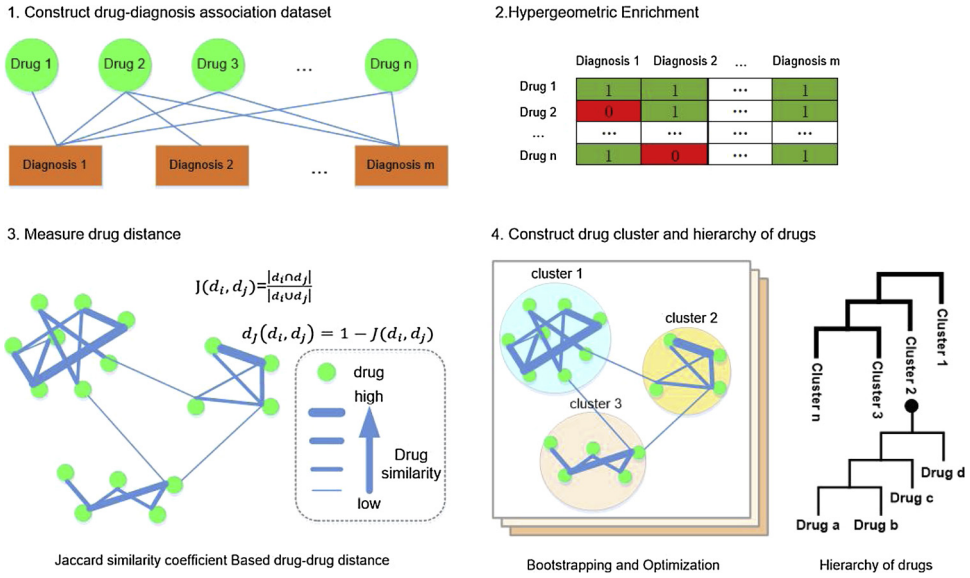

Fig. 1. Clinical drug–drug similarity analysis and clustering workflow. (1) Extract drug and diagnosis code from an EMR system. (2) Enrich drug-diagnosis association. (3) Measure drug–drug distance based on Jaccard similarity coefficient calculated from the enriched drug-diagnosis associations. (4) Cluster drugs and create hierarchical structures between and in clusters.

drug–drug similarity knowledge could not cover clinical drugs used in practice. Because in addition to the widely used Traditional Chinese Medicine, many of clinical drugs used in practice are compound drugs or bio-pharmaceuticals which were not included in current drug–drug similarity studies and drug classification systems.

Due to the tremendous growth of the promotion of EMR systems by many governments, more and more clinical data (e.g., demographics, diagnostic history, medication, laboratory test results, vital signs, discharge diagnosis) were accumulated electronically and ready for utilization [12]. The real-world clinical data which implies the clinical drug utilization knowledge, concomitantly provides drug–drug similarity based on drug and diagnostic relationships. In this study, we pose a hypothesis that the similarity of clinical drugs can be constructed from the drug-diagnosis association extracted from EMR data generated during routine practice in the hospital.

In this paper, we systematically design methods (as shown in Fig. 1) for constructing drug-diagnosis association, evaluate its statistical significance and measure clinical drug–drug distance based on it. These clinical drug–drug similarities are used to drive unsupervised drug clustering method. The hierarchical structures between clusters and in each cluster are further generated using two different approaches. Finally, we evaluate the drug–drug similarity matrix through comparing it with literature-based drug–drug similarity [1] and chemical similarity of drugs [8], discuss its potential translational utilization.

2. Materials and methods

2.1. Construct raw drug-diagnosis association dataset

An EMR system which implemented in a 2000-bedded hospital in China was used to construct the raw drug-diagnosis association database. For each in-patient encounter, its medication administration records and discharge diagnosis codes were extracted and linked with a unique encounter ID from EMR. Drugs and diagnosis codes which under the same in-patient encounter would generate a drug-diagnosis association pair. For instance, if the number of medications and discharge diagnosis codes belonging to an in-patient encounter is m and n respectively, there will be m x n pairs of drug diagnosis association.

A total of 812 554 medications and 339 269 discharge diagnosis codes were extracted from 53 922 encounters of 46 255 patients who were admitted to the hospital between January 1, 2016, and December 31, 2016. It generated 6 039 728 drug-diagnosis association pairs that contained 1210 distinct drugs and 6901 distinct diagnosis codes. Then, a 1210 × 6901 matrix was generated based on this raw dataset, and the individual items in the matrix were the count of the drug-diagnosis association pairs in the datasets.

2.2. Enrich statistically significant drug-diagnosis association

When a patient has multiple diagnosis codes, the associations between each of his drugs and diagnosis codes may just match by chance. Enrichment of associations between drugs and diagnosis codes was needed. To do so, a hypergeometric P value was calculated using the phyper function in R (version 3.4.0) for each drug-diagnosis pair which corresponds to the probability of having as many or more drug-diagnosis pairs conditioned on the full set of drug-diagnosis pairs. The Bonferroni correction was applied to control family-wise error rate for multiple testing. [13] After this enrichment, a matrix which contains Bonferroni-adjusted P value for each drug-diagnosis association was generated. All associations which P value was less than 0.05 were considered statistically significant and resulted in 20 297 statistically significant drug-diagnosis pairs (Supplemental data file named ENR-significant-drug-diagnosis-pair.xlsx). Top 10 associated drugs of five popular diseases from different disease classes were shown in supplemental Table S1. Most of drug-diagnosis pair is confirmed by the drug indications such as Acarbose Tablets for diabetes. Some of drugs do not have direct indications but always used combine with indicated drugs such as Palonosetron Injection was used to prevent nausea and vomiting that may occur within 24 h after receiving cancer chemotherapy.

2.3. Measure drug–drug distance

The drug–drug distance was measured based on the Jaccard similarity coefficient which performs on the significant drug-diagnosis association. To do so, the P values in the enriched matrix were converted to binary bits, where significant entries were set to 1 and nonsignificant entries were set to 0. The Jaccard similarity coefficient could be calculated as the proportion of bits for which both drugs with value 1 at same diagnosis code among those where at least one drug was 1(as shown in Formula (1)). And the drug–drug distance was calculated as Formula (2).

$$J(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|} \quad (1)$$

$$d_J(d_i, d_j) = 1 - J(d_i, d_j) \quad (2)$$

2.4. Construct drug clusters and hierarchical structures in each cluster

To cluster the 1210 drugs based on their pair-wise distances, the k-means clustering algorithm was selected because of its unsupervised feature and capability to handle large dataset [14]. We used *clusterboot* function from the *fpc* package in R which is an integrated function that computes the clustering by using interface function such as k-means and assures the cluster-wise stability of clustering of data [15]. As the value of k for the k-means algorithm is of great importance, we checked a broad range of conceivable numbers of k (from 10 to 100) corresponding to a large window around the commonly used value for k [16]. For each value of k, we performed 100 clusterboot bootstraps to prevent convergence to a local minimum. We used the Jaccard index to evaluate the goodness of clustering with a different value of k [17]. The optimal k was the k that maximized the mean Jaccard index. We performed 10,000 clusterboot bootstraps for the optimal k to guarantee the stability of drug cluster. To construct hierarchical structures in each cluster, we chose the hierarchical clustering which analysis on a set of dissimilarities matrix and aims to build a hierarchy of clusters. To do so, we used *hclust* function from the *stats* package in R.

2.5. Generate hierarchy of drug clusters based on similarity of diagnoses sets

For each drug cluster, there is a set of significantly associated diagnosis codes. We generated a kernel diagnosis code set for each drug cluster through selecting the top diagnosis codes which associated with more cluster drugs than half of the top 1 diagnosis code in this cluster. If more than 10 diagnosis codes meet this condition, only top 10 diagnosis codes are retained. We want to measure the distance between each drug cluster using the similarities of these kernel diagnosis codes sets. Fortunately, ICD-10 code as a hierarchy, taxonomy of the diseases contains the information to evaluate the semantic similarity between two diseases. Here the widely used Information content (IC) approach was used to assess the semantic similarity between two ICD-10 codes [18]. For the sake of notation, consider a set of diagnosis codes A, containing $\|A\|$ ICD-10 codes, where a denotes one of these ICD-10 codes. B is another set of ICD-10 codes and b is one ICD-10 code in B. Furthermore, c is defined as the least common ancestor (or super-concept) of two codes a and b in the ICD-10 hierarchy. The IC of a can be calculated by computing the count of taxonomical leaves of a concept's hyponym tree ($|leaves(a)|$) corresponding to its degree of generality and the number of taxonomical subsumers ($|subsumers(a)|$) representing its degree of concreteness [19].

$$IC(a) = -\log\left(\frac{\frac{|leaves(a)|}{|subsumers(a)|} + 1}{|leaves(root)| + 1}\right) \quad (3)$$

The similarity of two ICD codes is defined as:

$$CSim(a, b) = \frac{IC(a) + IC(b) - 2IC(c)}{IC(a) + IC(b)} \quad (4)$$

The similarity of two sets of ICD codes is defined as the average value of similarity scores of pairwise codes of two sets, given by:

$$SSim(A, B) = \frac{\sum_{a \in A, b \in B} CSim(a, b)}{|A| \bullet |B|} \quad (5)$$

The distance matrix among these diagnoses sets was used to generate a hierarchy of drug cluster by applying an agglomerative hierarchical clustering method called UPGMA.

Most of the statistical calculations and clustering were conducted under the R program language (version 3.4.0, the R script was available as supplement file drug-clustering.R) [21]. The similarities of diagnosis code sets were calculated based on a home-grown tool developed based on an ICD-10 knowledge base.

3. Results

3.1. Comparing with chemical similarity and literature-based similarity

To evaluate the clinical drug–drug similarity constructed in this study, we compared it with the chemical-based similarity and literature-based similarity of drugs. Firstly, we downloaded the US Food and Drug Administration–approved drug structures as an SDF file from DrugBank (accessed May 5, 2018; http://www.drugbank.ca/). DrugBank is a richly annotated resource that combines detailed drug data (i.e. chemical, pharmacological and pharmaceutical) with comprehensive drug target and drug action information [22]. In parallel, we downloaded the literature-based drug–drug similarity matrix generated by Brown et al [1]. A case-insensitive match between these drugs and DrugBank drugs resulted in 1489 overlapping drugs. Then we calculated extended connectivity fingerprints for each drug compound using the Morgan/circular method [23] and computed all pairwise chemical similarities using the Tanimoto similarity which has proven to be an appropriate choice for fingerprint-based similarity calculations [24]. Since all of the clinical data were extracted from a Chinese hospital, we removed drugs without significant association with diagnosis,carefully translated them into English and matched with the drug name mentioned above, finally resulting in 319 overlapping drugs in three drug

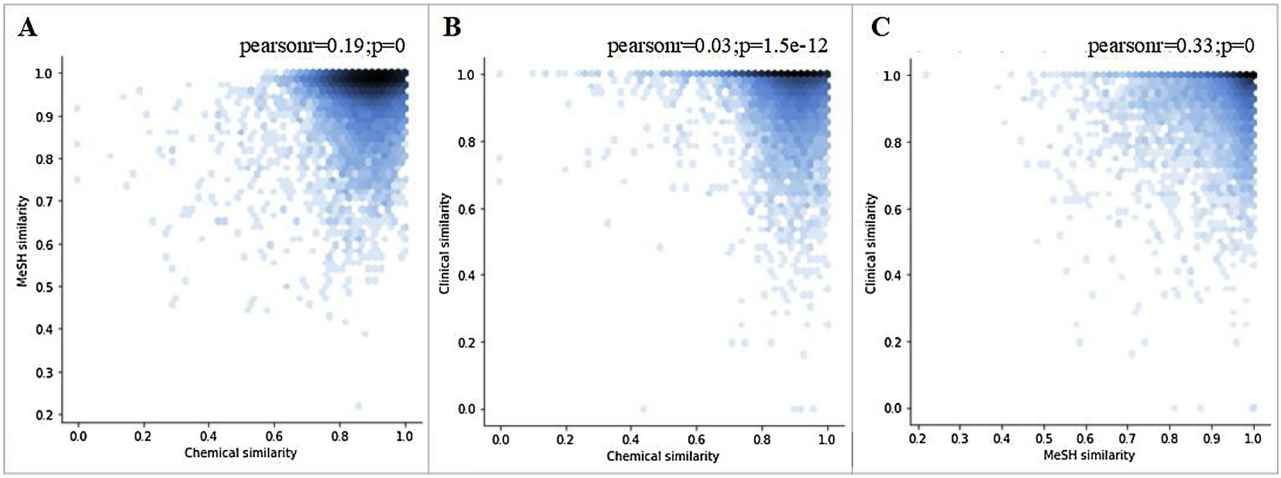

Fig. 2. Correlation analysis among different drug–drug similarity matrixes. A. The correlation of chemical similarity and MeSH similarity. B. The correlation of chemical similarity and clinical similarity. C. The correlation of MeSH similarity and clinical similarity.

list and it also showed the coverage of these similarities in clinical drugs. We compared three drug–drug similarity results in Fig. 2 (the value of 1 means dissimilarity).

In each part of the figure, drug similarities of 319 drugs calculated by two drug–drug similarity methods were binned in a hexbin with different color corresponding to the number of correlation at this bin. The Pearson's correlation coefficient and corresponding P value were calculated for each of the pairwise comparisons. All three similarity matrixes are positively correlated and the P values show the significant correlation among three drug–drug similarity methods. As there are many drugs which treat same clinical problems with very different chemical structure; for example, both *Miglitol Tablets* and *Glipizide Extended Release Tablets* are used for diabetes but with totally different chemical structure. On the other hand, there are also similar chemical structure drugs used under different clinical scenarios; for example, *Compound Betamethasone Injection* and *Dexamethasone Injection* with similar chemical structure but used in two different clinical departments for rheumatic diseases and cancer respectively. So, we do not expect a high correlation among these two drug similarity matrixes. While the clinical drug similarity shows a better correlation with the MeSH drug similarity.

3.2. Drug clusters

A total of 6 039 728 drug-diagnosis association pairs that contained 1210 distinct drugs and 6901 distinct diagnosis codes were extracted from an EMR dataset. After the enrichment, only 20 297 statistically significant drug-diagnosis pairs were retained. In the 100 clusterboot bootstraps, the mean Jaccard index got the highest value when k was set to 36 (Supplemental Fig. S1). After clustering the 1210 drugs, we obtained 36 mutually independent clusters. In the 36 clusters, there are two prominent clusters which contain more than 200 drugs (as shown in Fig. 3A). One is *Cluster1*, which contains 221 drugs, are not significantly associated with any diagnosis and not frequently used in clinical practice (The volume of drug used for each drug clusters was shown in the Supplemental Fig. S2). The other prominent cluster is *Cluster12*, which contains 276 drugs, only associated with a few diagnoses without many drugs share under it. Kernel diagnosis codes set for each drug clusters were summarized in supplemental Table S2. Using the similarities of these diagnosis codes set, the 36 drug clusters were organized in a hierarchy way as shown in Fig. 3B. Excepting the *Cluster1* which do not provide separated distances for each drug in it, all the drugs in other 35 drug clusters were plotted in 2D view in which each points represent a drug and distance among points reflect the relative distance among drugs. as shown in Fig. 3C which was based on the multidimensional scaling (MDS) of the drug–drug distance matrix.

Drugs in each drug cluster were further organized in a hierarchy way using the drug–drug distance. A website was developed to help the user exploring these clusters. Selecting a cluster N.O. on the webpage, the distribution of cluster drugs in all drugs, the hierarchical structure of drugs in the cluster, the similarity of each drug, associated diagnosis codes and ATC classes [20] of this cluster will be shown to users. The website can be accessed at http://kb4md.org:4000/drugcluster.

As shown in Fig. 3B, the 36 drug clusters can further be clustered in several groups. There are some isolated clusters such as *Cluster1*, *Cluster10*, *Cluster20*, *Cluster35* and *Cluster18* which are labeled with blue edges. *Cluster1* contains drugs without significant association with the diagnosis. The other four isolated clusters focus on specific clinical departments which are with distinctive clinical features. *Cluster10* and *Cluster20* contain drugs used in the department of gynecology and obstetrics. One for maternal care and the other for preterm newborn care. *Cluster35* contains drugs which treat mental disorders in the department of psychiatry. *Cluster18* contains drugs used in dermatology. The other 31 drug clusters can be organized in several hierarchical trees. The tree with red branches including *Cluster19*, *Cluster26*, *Cluster21*, *Cluster23*, *Cluster22* and *Cluster33* contain drugs treat a different type of cancers in the oncology department. The tree with green branches including *Cluster28*, *Cluster13* and *Cluster24* contain drugs deal with digestive system problems such as gastrointestinal problems, liver problems and pancreas problems. The golden branches including *Cluster9*, *Cluster32* and *Cluster27* contain drugs used when cerebral or spinal injury, fracture and related infection control. The purple branches of the tree including *Cluster2*, *Cluster30*, *Cluster8*, *Cluster34*, *Cluster12* and *Cluster11* are main branches focus on cardio-cerebral vascular diseases such as heart failure, atherosclerosis andcerebral infarction. The adjoining tree with pink branches contains drugs which are used to treat major endocrine and metabolic diseases such as diabetes, hyperlipidemia, hyperglyceridemia and their different complications in various systems such as kidney, neurological and circulatory systems. The last big tree with black branches whichcontains drugs mainly used in the respiratory systems. The clinical scenarios and detail kernel diagnosis codes set for each drug cluster were summarized supplemental Table S1.

To further underpin the relevance of the statistical significance and clinical drug–drug distance as well as the derived clusters. The 6 drug clusters which related to different cancers were highlighted in all drugs in supplemental Fig. S3A. The drugs with ATC code L (antineoplastic and immunomodulating agents) were also highlighted in supplemental Fig. S3B. As the figure shown, the 6 clusters not only covered most of antineoplastic agents but also provided many related drugs for cancer treatments. If we look closely at Cluster19 shown in Fig. S3C, there are several antineoplastic chemotherapy drugs such as *Cytarabine*, *Decitabine*, *Methotrexate*, *Mitoxantrone*, *Bortezomib*, *Arsenic Trioxide*, *Aclarubicin*, *Daunorubicin* and *Ephedrine*. At the same time, it also includes chemoprotectant (such as *Calcium Folinate* that helps prevent side effects of treatment or overdose with the medicine *Methotrexate* and similar medicines, *Recombinant Human Thrombopoietin* that helps correct the thrombocytopenia, *Diammonium Glycyrrhizinate Enteric-coated Capsule* that protects the liver function, *Omprazole Sodium* that helps to treat gastrointestinal bleeding during chemotherapy), anti-infective (such as *Aciclovir*, *Nystatin*, *Gentamycin* and *Fluconazole* that help the patient overcome the infective problems during the leukemia treatment) and immunopotentiator (such as *Placenta Polypeptide* and *Human Immunoglobulin*). In short, this drug cluster provided a relative complete medicine solution for the leukemia treatment not only the antineoplastic agents.

3.3. Alignment with the ATC classification system

To evaluate how this clinical drug–drug similarity results can be aligned to the ATC classification system, two heatmaps based on the drug–drug similarities by grouping drugs in ATC and drug clusters were shown in Fig. 4.

As we expected, when drugs grouped by 36 drug clusters generated in this study as shown in Fig. 4B, there are several dark square patches around the central diagonal for each cluster except two special drug clusters *Cluster1* and *Cluster12*. It means drugs in the same cluster with higher overall similarity. When grouping drugs using the ATC first level code as shown in Fig. 4A, although it is not so obviously there are also light dark square areas for well-defined ATC first levels such as A, B, C, J, L, M, N, P and R. The result supports the idea that drugs belonging to same ATC category are more likely to be similar. It also implies that there are many similar drugs classified in different ATC categories. The O and T which represent the drugs without ATC code assigned and the traditional Chinese medicine respectively are also highly related with other ATC groups. For example, the widely used TCM *Xiaoke Pills* are similar to other widely used anti-diabetic drugs such as *Metformin* and *Miglitol*.

4. Discussion

Different calculation methods of drug similarity have diverse application scenarios and advantages. For instance, chemical similarity

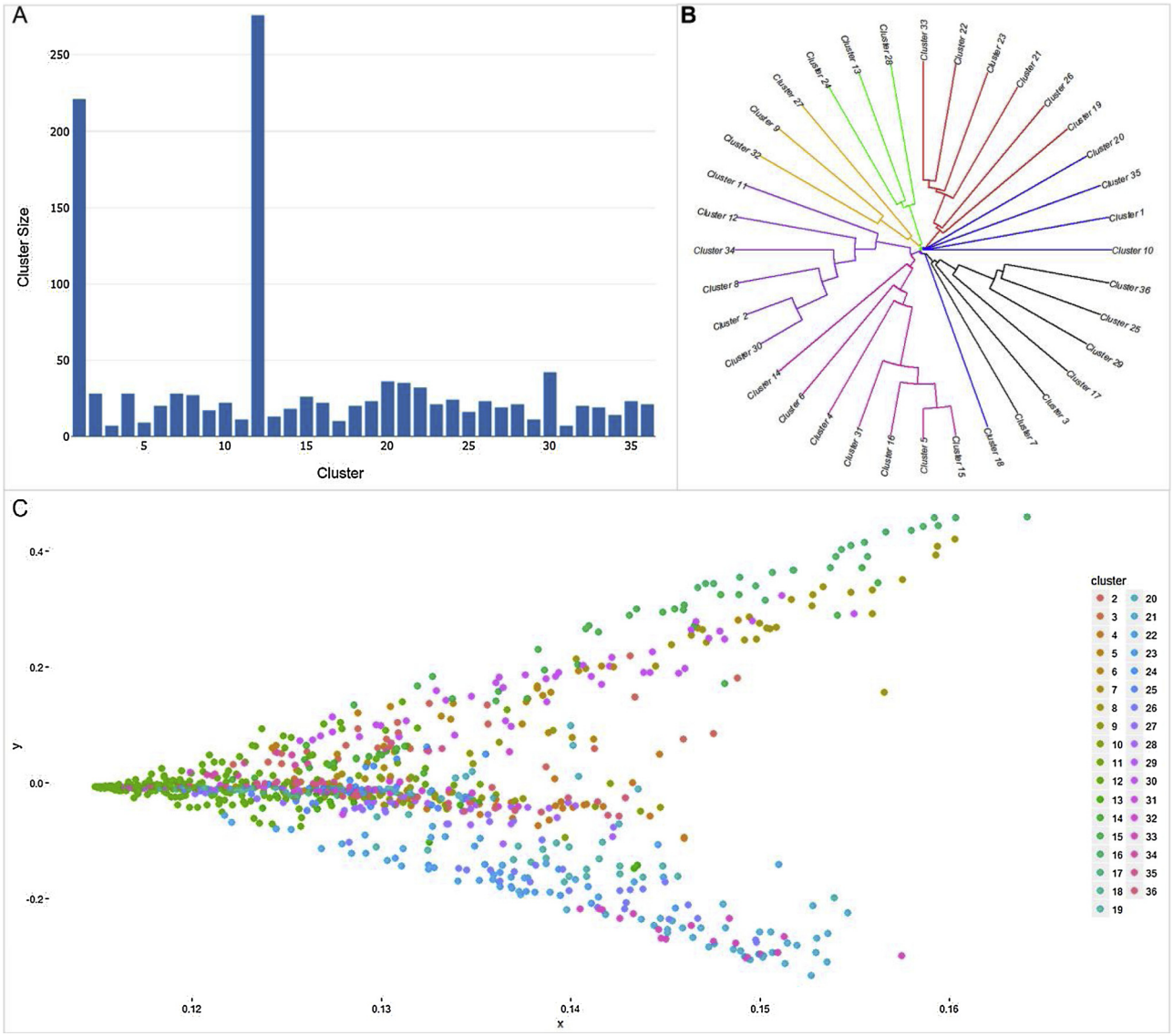

Fig. 3. Drug clusters based on clinical drug–drug similarity. A. The number of drugs in different drug clusters. B. Hierarchy of 36 clinical drug clusters. C. Multidimensional scaling of drug–drug distance matrix in 35 drug clusters.

plays an important role in predicting the properties of chemical compounds, finding underlying drug–drug interaction and especially conducting drug design studies. Nevertheless, only part of clinical drugs is single chemical substance, many of them are bio-pharmaceuticals or compound medicine and lack chemical structure data. MeSH similarity approach requires a given drug be represented in the biomedical

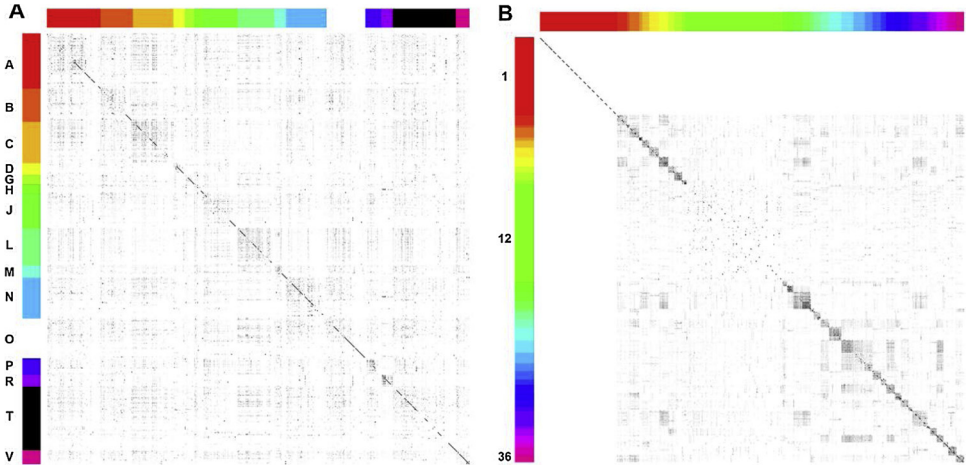

Fig. 4. Clinical drug similarity heatmap under different drug groups (the heatmap color scale from white to black and the darker color means higher drug similarity). A. Drug similarities grouped by ATC code. (O means drugs without an ATC code. T means traditional Chinese medicine.) B. Drug similarities grouped by clusters generated in this study.

literature and well-studied [1]. The clinical drug–drug similarity is calculated by extracting drug-diagnosis association from clinical data generated in the hospital and it includes all of the clinical drug used in practice. However, we do note that it has some limitations to measure the distance among infrequently used clinical drugs as the enriched drug-diagnosis associations are based on a probability calculation. While there are also many advantages of clinical drug–drug similarity to be used in clinical data utilizations.

4.1. Drug similarity and patient similarity

First, this clinical drug–drug similarity provided a way to quantitatively measure drugs similarity. Furthermore, the drug similarity can be used to quantitatively measure the similarity of medical therapy and further patient's similarity, which is an emerging concept in systems and precision medicine [25]. Patient similarity investigates distances between a variety of components of patient data and determines methods of clustering patients based on short distances between some of their characteristics. Among the similarome, which means a group of similar patients, index patient can be evaluated by further stratification guided by individual diagnoses, risk factors, medications, and so on. So far, there are many algorithms developed to estimate the different type of clinical data such as diagnosis, lab test result. But how to estimate the medication therapy similarity is not well established. This clinical drug–drug similarity method will pave a way for this kind of investigations.

4.2. Limitation

The limitations of this method should also be noted. As significant drug-diagnosis associations were calculated based on Bonferroni-adjusted hypergeometric P-value, the low-incidence diseases and orphan drugs will not be enriched. Therefore, the fact that there are thousands of unique and low-incidence diseases that do not have significantly associated drugs does not mean that there are no drugs for these conditions.

4.3. Mutual complementation

The TCM is widely used in China. As shown in Fig. 5, TCM spreads over various drug clusters. Traditionally, these drugs can only be classified using the TCM theories [26]. The chemical similarity and literature similarity approaches do not cover them and cannot be unified with the chemical drugs classification systems. However, in this proposed approach, these drugs were categorized along with other drugs at the meanwhile.

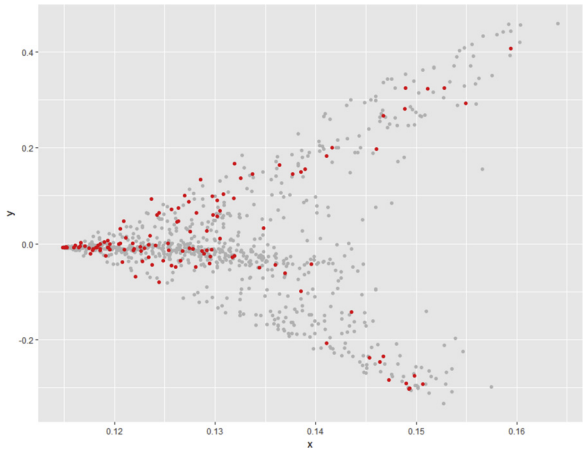

Fig. 5. The Traditional Chinese Medicine distribution in all drugs (The red spot is TCM) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Meanwhile, there are many infrequently used clinical drugs were not fully manifested in this approach and some of them were well aligned in other methods. The distances information about these drugs in other drug–drug similarity systems can be used to complement the clinical drug–drug similarity and vice versa. In a summary, three different approaches with different advantages were mutual complementation.

5. Conclusion

A measurement of drug–drug similarity covers all clinical drugs in a hospital were proposed in this study. We compared this method with other drug similarity results and discussed the advantages of different methods. This clinical drug–drug similarity also provided a quantitative measurement foundation to calculate the drug similarities and therapy similarities.

Author contributions

XZ, JZ, and HL wrote the manuscript. HL and XZ designed the research. ZH and WC provided expertise explanation of the drug cluster. XL and HD collected the EMR data. XZ, JZ and HL performed the research. XZ, ZH and HL analyzed the data. HL developed the website.

Conflict of interest

The authors declared no conflict of interest.

Summary table

What was already known on the topic:

- Drug similarity can be measured based on different drug features such as chemical structure
- Drug similarity can be measured based on the co-occurring of drug and MeSH in literature
- All these drug similarity results can not cover the clinical drugs used in practice

What this study added to our knowledge:

- Clinical drug similarity can be measured from clinical datasets
- Clinical drug–drug similarity correlated with chemical similarity and literature-based drug similarity
- Clinical drug similarity can be aligned with current drug classification systems such as ATC

Acknowledgments

We acknowledge the support of the Shanxi Dayi Hospital (Shanxi, China) for supplying the anonymized clinical data. This work was inspired by the work of Dr. Chirag J Patel in MeSHDD. H.L. was supported by the National Natural Science Foundation of China (81871456) and National Key R&D Program of China (2016YFC0901905).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2019.02.003.

References

[1] A.S. Brown, C.J. Patel, MeSHDD: literature-based drug–drug similarity for drug repositioning, J. Am. Med. Inform. Assoc. 24 (2017) 614–618.
[2] P. Zhang, F. Wang, J. Hu, Toward drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity, AMIA Annu. Symp. Proc. 2014 (2014) 1258–1267.
[3] H. Luo, J. Wang, M. Li, et al., Drug repositioning based on comprehensive similarity measures and Bi-random walk algorithm, Bioinformatics 2016 (32) (2016) 2664–2671.
[4] R. Ferdousi, R. Safdari, Y. Omidi, Computational prediction of drug–drug interactions based on drugs functional similarities, J. Biomed. Inform. 70 (2017) 54–64.
[5] S. Dhanya, F. Shobeir, G. Lise, A probabilistic approach for collective similarity-based drug–drug interaction prediction, Bioinformatics 32 (2016) 3175–3182.
[6] M. Campillos, M. Kuhn, A.C. Gavin, et al., Drug target identification using side-effect similarity, Science 321 (2008) 263–266.
[7] E. Lounkine, M. Keiser, et al., Large-scale prediction and testing of drug activity on side-effect targets, Nature 486 (2012) 361–367.
[8] P. Zhang, F. Wang, J.Y. Hu, et al., Toward personalized medicine: leveraging patient similarity and drug similarity analytics, AMIA Jt. Summits Transl. Sci. Proc. 2014 (2014) 132–136.
[9] K. Cha, M.S. Kim, K. Oh, Drug similarity search based on combined signatures in gene expression profiles, Healthc. Inform. Res. 20 (2014) 52–60.
[10] N.P. Tatonetti, P.P. Ye, R. Daneshjou, et al., Data-driven prediction of drug effects and interactions, Sci. Transl. Med. 4 (2012) 125ra31.
[11] R. Sawada, H. Iwata, S. Mizutani, et al., Target-based drug repositioning using large-scale chemical-protein interactome data, J. Chem. Inf. Model. 55 (2015) 2717–2730.
[12] H. Sun, K. Depraetere, J. De Roo, et al., Semantic processing of EHR data for clinical research, J. Biomed. Inform. 58 (2015) 247–259.
[13] R.A. Armstrong, When to use the Bonferroni correction, Ophthalmic Physiol. Opt. 34 (2014) 502–508.
[14] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666.
[15] C. Hennig, fpc: Flexible Procedures for Clustering. https://CRAN.R-project.org/package=fpc. (Accessed 22 June 2017).
[16] D. Steinley, M.J. Brusco, Choosing the number of clusters in K-means clustering, Psychol. Methods 16 (2011) 285–297.
[17] C. Hennig, Cluster-wise assessment of cluster stability, Comput. Stat. Data Anal. 52 (2007) 258–271.
[18] S. Harispe, D. Sanchez, D. Ranwez, et al., A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain, J. Biomed. Inform. 48 (2014) 38–53.
[19] D. Sánchez, M. Batet, D. Isern, Ontology-based information content computation, Knowl. Based Syst. 24 (2010) 297–303.
[20] WHO Collaborating Centre for Drug Statistics Methodology. ATC: Structure and principles. https://www.whocc.no/atc/structure_and_principles. (Accessed 22 September 2017).
[21] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, J. Comput. Graph. Stat. 5 (1996) 299–314.
[22] D.S. Wishart, C. Knox, A.C. Guo, et al., DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906.
[23] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.
[24] D. Bajusz, A. Racz, K. Heberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminform. 7 (2015) 20.
[25] S.A. Brown, Patient similarity: emerging concepts in systems and precision medicine, Front. Physiol. 7 (2016) 561.
[26] X. Liu, Q. Wang, G. Song, et al., The classification and application of toxic Chinese material medica, Phytother. Res. 28 (2014) 334–347.