RESEARCH

Open Access

Clinical connectivity map for drug repurposing: using laboratory results to bridge drugs and diseases

Qianlong Wen[1†], Ruoqi Liu[2†] and Ping Zhang[2,3*]

Abstract

Background: Drug repurposing, the process of identifying additional therapeutic uses for existing drugs, has attracted increasing attention from both the pharmaceutical industry and the research community. Many existing computational drug repurposing methods rely on preclinical data (e.g., chemical structures, drug targets), resulting in translational problems for clinical trials.

Results: In this study, we propose a novel framework based on clinical connectivity mapping for drug repurposing to analyze therapeutic effects of drugs on diseases. We firstly establish clinical drug effect vectors (i.e., drug-laboratory results associations) by applying a continuous self-controlled case series model on a longitudinal electronic health record data, then establish clinical disease sign vectors (i.e., disease-laboratory results associations) by applying a Wilcoxon rank sum test on a large-scale national survey data. Eventually, a repurposing possibility score for each drug-disease pair is computed by applying a dot product-based scoring function on clinical disease sign vectors and clinical drug effect vectors. During the experiment, we comprehensively evaluate 392 drugs for 6 important chronic diseases (include asthma, coronary heart disease, congestive heart failure, heart attack, type 2 diabetes, and stroke). The experiment results not only reflect known associations between diseases and drugs, but also include some hidden drug-disease associations. The code for this paper is available at: https://github.com/HoytWen/CCMDR

Conclusions: The proposed clinical connectivity map framework uses laboratory results found from electronic clinical information to bridge drugs and diseases, which make their relations explainable and has better translational power than existing computational methods. Experimental results demonstrate the effectiveness of our proposed framework, further case analysis also proves our method can be used to repurposing existing drugs opportunities.

Keywords: Drug repurposing, Connectivity map, Electronic health record, National Health and Nutrition Examination Survey

*Correspondence: mail.pingzhang@gmail.com
†Qianlong Wen and Ruoqi Liu have contributed equally to this work.
[3] Department of Biomedical Informatics, The Ohio State University, 1800 Cannon Drive, Columbus, Ohio 43210, USA
Full list of author information is available at the end of the article

Background

Traditional *de novo* drug discovery is a long and complicated process [1, 2], which usually takes more than 15 years [3], and costs 800 million to 1 billion US dollars [4] to develop a new drug. Drug repurposing, investigation of potential additional uses for existing drugs, is

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 2 of 11

becoming an appealing research field given its potential in lowering overall costs and shortening drug development timelines [5].

There has been a surge of computational methods proposed for drug repurposing in recent years, which can be roughly classified into two categories based on different data sources: preclinical data-based and clinical data-based. Preclinical data-based methods often build machine learning models based on preclinical data, such as drug chemical structure, protein targets and gene expression information, to identify potential drug-disease associations. For example, Keiser et al. [6] use drug structural similarity as the measurements to find the drugs with similar effects. Lamb et al. [7, 8] raise the connectivity map (CMap) approach for drug repurposing by using gene expression data, which is based on molecular activity. Luo et al. [9] develop a server named DPDR-CPI which predicts the new indications of existing drugs by analyzing the chemical-protein interactome (CPI) profile. Some researchers also tried to construct computational frameworks that integrated several kinds of data sources and even disease similarity measurement profiles to make better predictions. PreDR model proposed by Wang et al. [10] integrated drug structure, drug target, side-effects and disease phenotype data to find the novel drug indications. Zhang et al. The similarity constrained matrix factorization method raised by [11] take known drug-disease associations, drug features and disease semantic information as input to predict drug-disease association. However, all of these methods rely heavily on preclinical information to make predictions. This will cause a large translation gap when we apply the drugs on humans. It is estimated that of all compounds effective in cell assays, only 30% of them could work in animals and only 5% of them could work in humans [12].

Compared with preclinical data, clinical data provide more applicable and reliable data sources for drug repurposing as clinical information (e.g., laboratory test results) because it records direct read-outs drug effects on patients, so there is no need to consider about the translational problems. Many computational frameworks based on clinical information has been raised due to the large amount of available electronica clinical data.
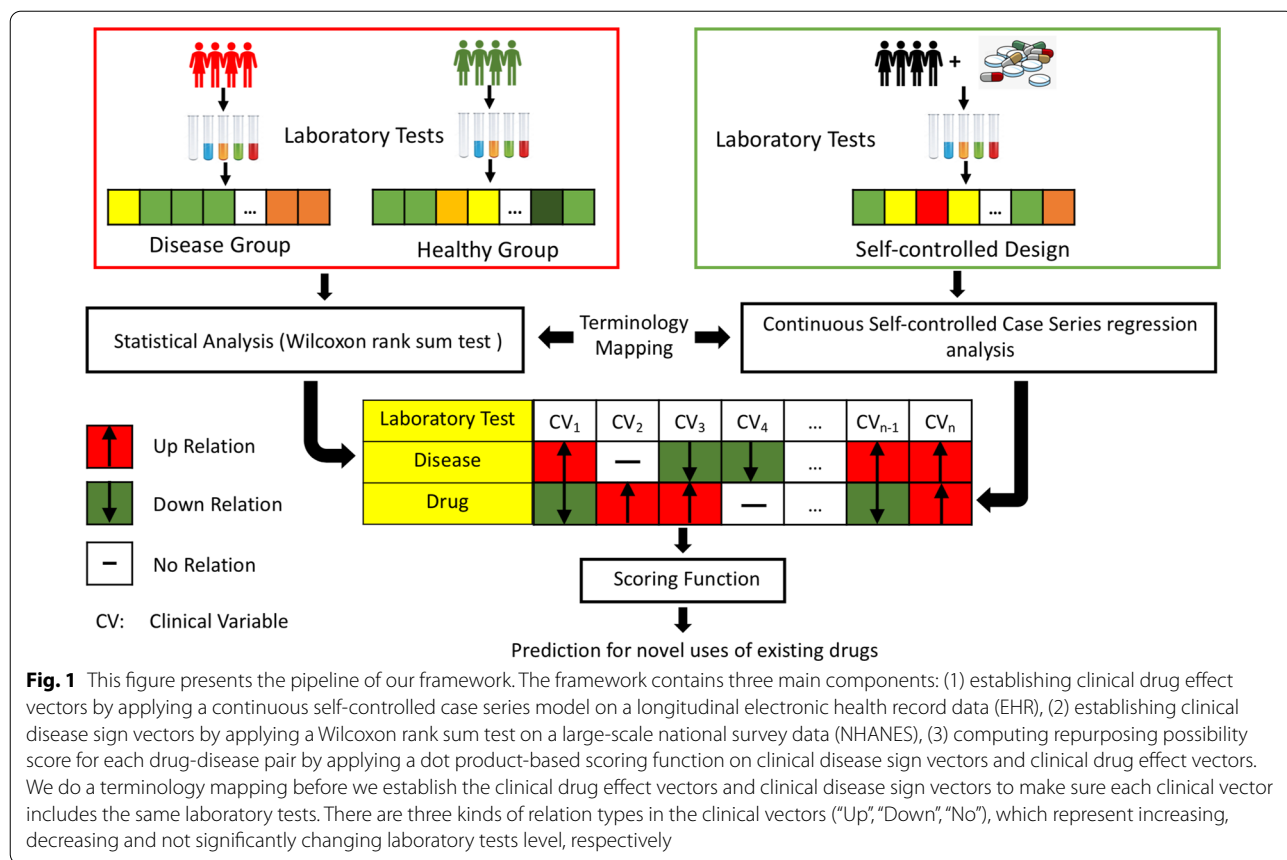
Jung et al. [13] find the connection between drugs and diseases in clinical diagnose notes by literature mining, but it does not include any other structured data, like laboratory test results. Jang et al. [14] propose a framework that use laboratory test results to reflect the influence of drugs and diseases on human physiological activities, and the method they use to establish drug effects is counting co-occurrence between drug and laboratory tests. However, it is not efficient enough to dig the hidden relation between drugs and laboratory tests, especially when we

have a large dataset and include many laboratory and existing drugs in our experiment. Kuang et al. [15] and Ghalwash et al. [16] raised more advanced methods to compute the influence of drugs on laboratory tests, however, they reflect the effect of drugs on single laboratory (e.g., blood sugar level), which it is not enough to represent the state of the complex human system. It would be more efficient and accurate if we build an electronic clinical information-based drug repurposing framework and implement it by more efficient statistical analysis methods designed for large datasets. During this process, we will include as many laboratory tests as we can in our experiment to completely represent the state of human biological system. The idea of CMap raised by Lamb et al. [7, 8] which uses gene expression values to bridge drugs and diseases, directly inspires us to formulate and leverage all the laboratory tests involved in our experiment to build associations between drugs and diseases from clinical perspective.

In this paper, we propose a clinical connectivity map framework for drug repurposing (CCMDR) by leveraging laboratory tests to analyze the influence of drugs and diseases on the human biological system. The overall framework is illustrated in Fig. 1. Experimental results show that our method can not only retrieve the known drug-disease associations in high accuracy but also can find potential indications, which can be verified from medical literature. Moreover, the associations between the predicted drug-disease can be clearly and vividly represented via the corresponding complementarity between laboratory tests of drug effect vectors and disease sign vectors, which make our results more explainable. Thus, the evaluation performance and explainability show the potential that our method could be used in future drug repurposing tasks.

In brief, the contribution of the paper can be summarized as below:

- We propose a clinical connectivity mapping framework for drug repurposing. The new framework solely based on the clinical patient data, thus with less translational problems.
- We evaluate our framework for 392 drugs on 6 important chronic diseases (include asthma, coronary heart disease, congestive heart failure, heart attack, type 2 diabetes, and stroke). Experimental results show that our method achieves high accuracy in retrieving the known indications of drugs.
- We study the predicted drug repurposing candidates via the corresponding complementarity between laboratory tests of drug effect vectors and disease sign vectors. Case studies with literature support show

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 3 of 11

Fig. 1 This figure presents the pipeline of our framework. The framework contains three main components: (1) establishing clinical drug effect vectors by applying a continuous self-controlled case series model on a longitudinal electronic health record data (EHR), (2) establishing clinical disease sign vectors by applying a Wilcoxon rank sum test on a large-scale national survey data (NHANES), (3) computing repurposing possibility score for each drug-disease pair by applying a dot product-based scoring function on clinical disease sign vectors and clinical drug effect vectors. We do a terminology mapping before we establish the clinical drug effect vectors and clinical disease sign vectors to make sure each clinical vector includes the same laboratory tests. There are three kinds of relation types in the clinical vectors ("Up", "Down", "No"), which represent increasing, decreasing and not significantly changing laboratory tests level, respectively

the potential of our method to discover previously unknown indications of existing drugs.

Methods

Dataset and data preprocess

We use the questionnaire and laboratory results from the National Health and Nutrition Examination Survey (NHANES) [17] to establish the clinical disease sign vectors. According to the questionnaire survey (e.g., "Has been diagnosed with type 2 diabetes?"), individual samples are divided into disease group (who answered "yes") and healthy group (who answered "no"). Next, we perform the statistical analysis to identify those disease-related clinical variables from collected laboratory results in NHANES data. We extract 87,464 individual samples, 986 numerical clinical variables and more than 30 disease conditions from NHANES data range from 1999 to 2016. Here, we only consider the disease conditions with more than 1000 individual samples, which results in 6 unique diseases (i.e., asthma, coronary heart disease, congestive heart failure, heart attack, type 2 diabetes and stroke).

We use the prescription and laboratory result histories of patients in a proprietary deidentified Electronic Health Record (EHR) to establish the clinical drug effect vectors. We transform the prescription records of patients into matrixes based on medication use situations. To study the associations between prescribed drugs and laboratory results, we apply a continuous self-controlled case series model [15] to analyze the effects of a drug on the laboratory results. We only consider patients with complete records (i.e., having both prescription and its corresponding laboratory results), which results in 91,934 patients, 1344 kinds of treatments and 65 kinds of laboratory results. After excluding those prescriptions with less than 1000 patients, we obtain 392 unique prescribed drugs.

We bridge the drug and disease using the laboratory results obtained from each side. Since the laboratory results are from different data resources (i.e., national survey data and electronic health records), we need to standardize those laboratory results for further analysis. The laboratory results that appear in both datasets are included and mapped to a standard list with consistent names. Also, the non-numerical laboratory results are excluded. Finally, we obtain 35 laboratory results considered as clinical variables. The full list of the 35 clinical variables can be found in Additional

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 4 of 11

file [5]: Table S1. Our inference of a drug-disease pair is based on the complementary and adverse effects that each drug candidate and disease condition has on the 35 clinical variables.

Clinical disease sign vector

We extract 6 disease conditions and 35 clinical variables from NHANES after preprocessing to establish the clinical disease sign vector. The dimension of each disease sign vector is 1×35. There are three types of relations between a disease and clinical vectors (i.e., "Up", "Down" and "No"), which represents increasing, decreasing and not significantly changing of laboratory results level, respectively. As mentioned above, the combined data is divided into disease group and control group according to the questionnaire data, we apply Wilcoxon rank sum test (a.k.a., Mann Whitney U test) on two groups to calculate the p-value for each clinical variable. Certain p-value cut-off is used to examine whether the value change is significant or not [18]. In our work, the p-value threshold is set to be 0.05. Only the clinical variables satisfy the condition that p-values are less than 0.05 can be regarded as significant clinical variables concerning the disease. We consult the Mann–Whitney table of $\alpha = 0.05$. If the smaller value of U_1 and U_2 is larger than the value given in the table, the null hypothesis is true otherwise false. Then we assign relation direction to this clinical variable by comparing the average clinical variable value of the disease group and control group. Up relation ("↑") indicates a significant value increase in the disease group compared with the control group, while down relation ("↓") means the laboratory value of the disease group is significantly lower than that of the control group, no relation ("-") indicate the laboratory result level will not be significantly influenced by the disease.

Clinical drug effect vector

To establish clinical drug effect vectors, we extract 392 drugs and 35 clinical variables from EHR data. The clinical variables used here are the same as ones in establishing the disease sign vectors. So, the dimension of each drug effect vector is 1×35. We need to consider the prescription records of patients and their corresponding laboratory results records simultaneously, and the EHR dataset we use is a large dataset that includes millions of records. So, we need to find a way to analyze the high-dimensional longitudinal data. In our work, we adopt the continuous self-controlled case Series (CSCCS) model proposed by Kuang et al. [15], it is a lasso regression analysis model designed to do the data analytical work for EHR dataset.

Assuming there are N patients with a specific kind of clinical variable measurement and M kinds of drugs in EHR dataset. Continuous variable y_{ij}, where $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, J_i\}$, indicates the value of j_{th} clinical variable measurement taken among a total number of J_i measurements for the i_{th} patient, while binary variable x_{ijm}, where $i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, J_i\}$, $m \in \{1, 2, \ldots, M\}$, are used to indicated the drug whether i_{th} patient are exposed to the m_{th} drug when the j_{th} clinical variable measurement is taken. 0 represents no and 1 represents yes.

y_{ij} is regard as the output variables when we fit the structured data into the linear regression model, so we have:

$$y_{ij}|\boldsymbol{x}_{ij} = \alpha_i + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_M \end{bmatrix}^\top, \quad (1)$$
$$\boldsymbol{x}_{ij} = \begin{bmatrix} x_{ij1} & x_{ij2} & \cdots & x_{ijM} \end{bmatrix}^\top$$

α_i in Eq. (1) represents the average baseline level of y_{ij} on i_{th} patient. That means it is independent of the date the measurement was taken and drugs the patient used when the measurement was taken. Each patient has an individual baseline value. ϵ_{ij} here is an independent and identically distributed Gaussian noises with zero means and fixed but unknown variance σ^2. Then the linear model can be easily converted to a least square problem as follows:

$$\arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{1}{2} \left\| \boldsymbol{y} - \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 \quad (2)$$

where

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_N \end{bmatrix}^\top, \boldsymbol{Z} = \text{diag}\left(\boldsymbol{1}_1, \cdots, \boldsymbol{1}_N\right),$$
$$\boldsymbol{y} = \begin{bmatrix} y_{11} & \cdots & y_{1J_1} & \cdots & y_{N1} & \cdots & y_{NJ_N} \end{bmatrix}^\top,$$
$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1J_1} & \cdots & \boldsymbol{x}_{N1} & \cdots & \boldsymbol{x}_{NJ_N} \end{bmatrix}^\top$$

where $\boldsymbol{Z}$ is a block diagonal matrix and $\boldsymbol{1}_i$ is a $J_i \times 1$ vector in which all the components are 1. By solving this problem, we can get the optimized parameter $\boldsymbol{\beta}$, which is also the interest of our task. $\boldsymbol{\beta}$ is a $1 \times M$ parameter vector, parameter $\boldsymbol{\beta}_m$ in $\boldsymbol{\beta}$ indicates the effect of m_{th} drug on the output variable $\boldsymbol{y}$. The optimized parameter we get with the CSCCS model is numerical. Positive and negative parameters in this vector represent the corresponding drugs that may increase and decrease the level of output variable respectively, while 0 indicates the corresponding drugs do not influence it. In the CSCCS model, parameter $\boldsymbol{\alpha}$ is regarded as a nuisance parameter, our interest is parameter $\boldsymbol{\beta}$ so we do not need to care the value of $\boldsymbol{\alpha}$. To eliminate the effect of $\boldsymbol{\alpha}$, [15] consider:

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 5 of 11

$$\frac{\partial \mathcal{L}(\alpha, \beta)}{\partial \alpha} = 0 \Rightarrow \alpha = \left(Z^\top Z\right)^{-1} Z^\top (y - X\beta) = \overline{y} - \overline{X}\beta \tag{3}$$

Where $\overline{y}$ is a $N \times 1$ vector which includes the average value of clinical value among N patients, $\overline{y}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij}$. $\overline{X}$ is a $N \times M$ matrix and $\overline{X}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} x_{ij}$. So, the expression of CSCCS model below, which is free of α, is derived by substituting Eq. (3) into Eq. (2):

$$\arg\min_\beta \frac{1}{2} \|y - Z\overline{y} - (X - Z\overline{X})\beta\|_2^2 \tag{4}$$

When we apply the CSCCS model on the high-dimensional longitudinal EHR data, we will add a L_1 penalty term because there is an assumption that the level of clinical variables will only be significantly influenced by a small portion of drugs. The L_1 penalization drives most components of β to zero or closed to zero [19]. In other words, we simply want to know the drugs which are most correlated to the level change of clinical variables. So the final expression of the CSCCS model we apply to this problem is:

$$\arg\min_\beta \frac{1}{2} \|y - Z\overline{y} - (X - Z\overline{X})\beta\|_2^2 + \lambda\|\beta\|_1 \tag{5}$$

where $\lambda > 0$, λ decides the sparsity of optimized result so we need to tune this parameter to get a final result with proper sparsity level.

In order to further filter out the drugs which do not have a significant effect on clinical variables, our implementation also returns the p-value of each component in β. We apply the same p-value cut-off strategy on the optimized result. Parameters with a p-value greater than 0.05 in β are regarded as insignificant effect and we assume their corresponding drugs are uncorrelated with clinical variable level change. The significant effects can be divided into increasing or decreasing effect based on the coefficient value is positive or negative. Then we assign each drug-clinical variable pair up ("↑"), down ("↓") and no ("-") relation type just like clinical disease sign vectors.

Scoring function

After we establish the clinical vectors for each drug-disease pair, we need to define a scoring function to calculate the repurposing possibility score for each drug-disease pair. The inference for each drug-disease pair is based on complementary and adverse effects. Specifically, complementary effect refers to the opposed relation type between a clinical disease sign vector and clinical drug effect vector on the same clinical variables, while adverse effect refers to the same relation type between a clinical disease sign vector and clinical drug effect vector

on the same clinical variables. The complementary relation direction between the two vectors will increase the final repurposing possibility score of a drug-disease pair while adverse relation direction will decrease it. Here, we use a dot product-based scoring function to consider both complementary and adverse effects of a drug candidate on a disease. The scoring function can be written as follow:
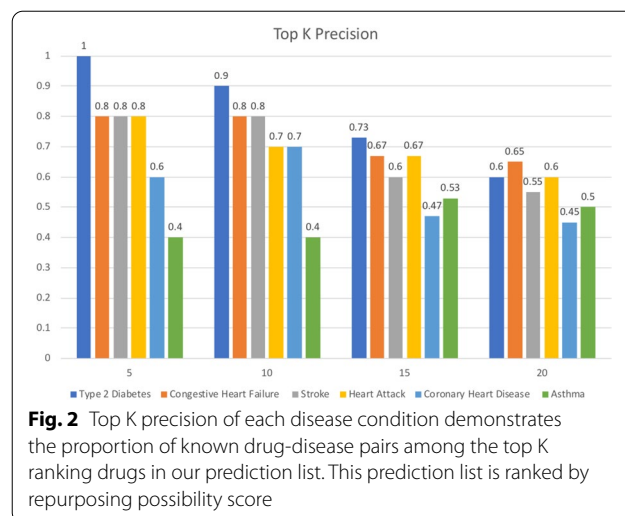
$$\text{TS}_{\text{disease-drug}} = -\text{CV}_{\text{Drug}} \cdot \text{CV}_{\text{Disease}} \tag{6}$$
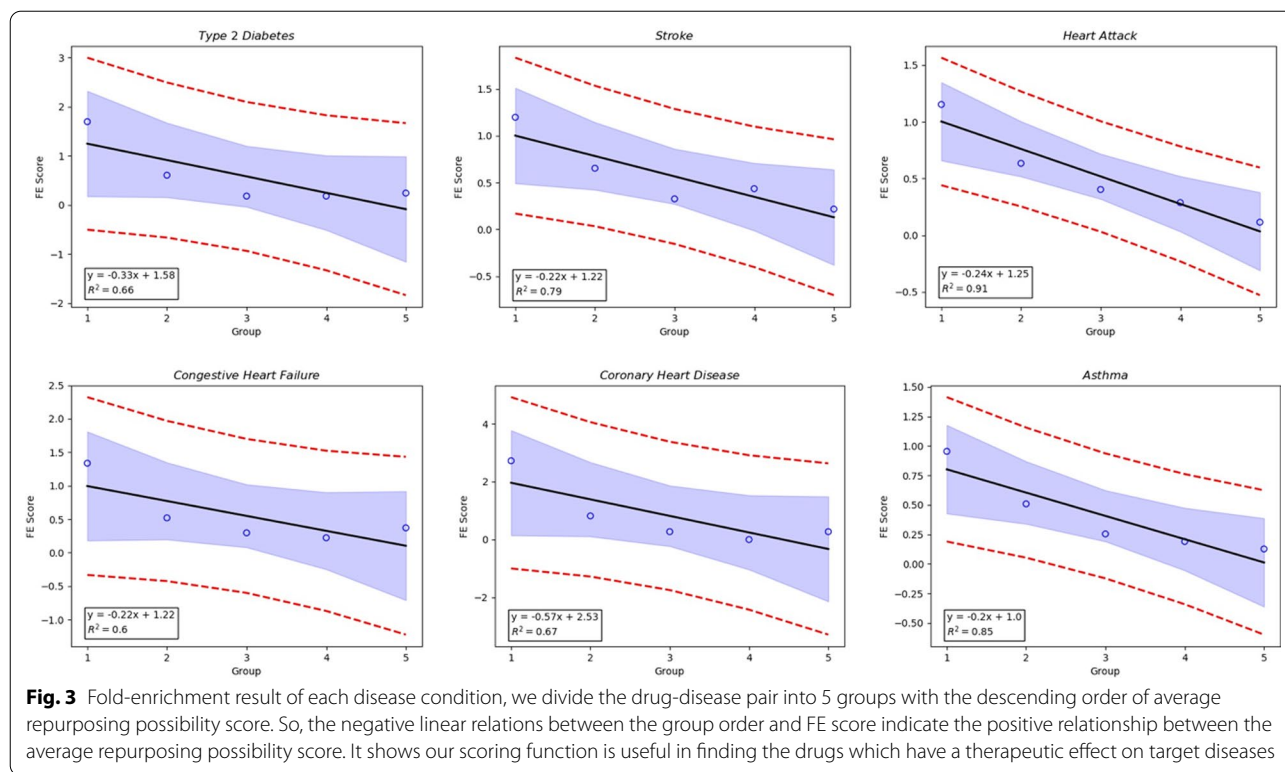
where CV_{Drug} is the clinical drug effect vector, and $\text{CV}_{\text{Disease}}$ is the clinical disease sign vector. We transform the 3 kinds of relation type in clinical vectors ("↑", "↓" and "−") into numerical values (1.0, −1.0, 0.0) for the convenience of calculation. To rank the drugs in descending order and emphasize the most powerful drug candidates predicted by our model, we add a minus sign before the product of the two vectors. So, the positive result calculated by this scoring function means there are more complementary relation directions than adverse relation directions between a drug candidate and a disease, while negative results indicate more adverse relation directions between this pair.

Results and discussion

Evaluation metrics

After we calculate the repurposing possibility score of each drug-disease pair, we need to prove that the score is qualified enough to serve as a metric to show whether a drug candidate is likely to be the potential treatment or not. The final drug candidate list is sorted by the repurposing possibility score in a descending order for the convenience of validation. The validation data we use comes from Side Effect Resource(SIDER) [20]. It contains drugs with indications or side-effects for many kinds of disease

Fig. 2 Top K precision of each disease condition demonstrates the proportion of known drug-disease pairs among the top K ranking drugs in our prediction list. This prediction list is ranked by repurposing possibility score

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 6 of 11

Fig. 3 Fold-enrichment result of each disease condition, we divide the drug-disease pair into 5 groups with the descending order of average repurposing possibility score. So, the negative linear relations between the group order and FE score indicate the positive relationship between the average repurposing possibility score. It shows our scoring function is useful in finding the drugs which have a therapeutic effect on target diseases

conditions. We take it for ground truth to testify whether our method can retrieve known indications of drugs. The hypothesis is that drug candidates with higher repurposing possibility score are more likely to be the treatment of the disease, which means most of the top-ranking drugs can be found in the drugs with indication and most of the bottom ranking drugs can be found in the drugs with side-effects provided by SIDER. In this case, those drugs can not be found in the validation data but still predicted with high repurposing possibility score by our model could be served as a potential treatment of the disease. So we need to use some evaluation metrics to test whether the known drug-diseases pairs are enriched at the top of our prediction list. We will use two kinds of evaluation metrics to validate our prediction.

Precision at K
The First kind of evaluation metric is precision at K. The top K precision value is the ratio of known treatment for a disease among the top K drug candidates for the disease predicted by our framework.

For each disease, we rank the drugs using the calculated repurposing possibility score. Then we compute the precision at K ($K \in \{5, 10, 15, 20\}$) of each disease using the top-ranked K drugs (e.g., precision at 10 corresponds to the proportion of correct retrieved drugs among the top 10 ranked drugs).

Fold-enrichment test
Another evaluation metric to access whether our repurposing possibility score is correlated with the likelihood that disease-drug pair occurs or not is the fold-enrichment (FE) test. FE score can be defined by the following formula:

$$\text{FE Score} = \frac{(n/m)}{(N/M)} \tag{7}$$

where M is the number of all the mapped drugs and N is the number of drugs in the gold-standard dataset corresponding to each kind of disease condition. We will divide all the mapped drugs evenly into several groups according to their repurposing possibility score. So, m is the total number of drugs in one group and n is the number of drugs involved in the gold-standard dataset within the group. FE test can demonstrate the enrichment of known disease-drug pair (we assume the drug-disease pairs in SIDER is ground truth) within different score ranges. Our prediction can be proved to be reasonable if the FE test score is positively correlated with the repurposing possibility score. There are 392 drug-disease pairs for each kind of disease condition in our experiment, and all of them are ranked by repurposing possibility score and binned into groups of 80 pairs (the last group contains 72 drug-disease pairs). The scoring function is reasonable if the FE score is decreasing with the ascending

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 7 of 11

Table 1 This table presents the selected previously unknown drug-disease pairs predicted by our method, we just show the clinical variables that contribute to the final repurposing possibility of each drug-disease pair in the table

Laboratory result		Alkaline phosphatase	Cholesterol	Glucose	HDL	LDL	Triglycerides
Disease	Type 2 diabetes	↓	↑	↑	↓	↑	↑
Drug	Clopidogrel Hydrogen sulfate	–	↓	–	–	↓	–
Disease	Type 2 Diabetes	↓	↑	↑	↓	↑	↑
Drug	Doxycycline hyclate	–	–	↓	–	–	–
Disease	Coronary heart disease	↑	↑	↑	↓	↑	↑
Drug	Alendronate sodium	↓	–	–	↑	–	–
Disease	Congestive heart failure	↑	↑	↑	↓	↑	↑
Drug	Alendronate sodium	↓	–	–	↑	–	–
Disease	Heart attack	↑	↑	↑	↓	↑	↑
Drug	Alendronate sodium	↓	–	–	↑	–	–

order of the 5 groups because the average repurposing possibility score of each group is decreasing in that order.

Established disease and drug vectors

In our experiment, we first establish all of the clinical diseases and drug vectors. All the clinical disease sign vectors are represented in Additional file 1: Table S2, and all the clinical drug effect vectors are represented in Additional file 2: Table S3.

Then, we calculate the repurposing possibility score of 392 kinds of drugs on six disease conditions (asthma, coronary heart disease, congestive heart failure, heart attack, type 2 diabetes and stroke). The repurposing possibility score of each drug-disease pair is listed in Additional file 4: Table S4. We also transform the table into a heat map Fig. 4 to vividly present the repurposing possibility score. Due to page limitation, we just present the drugs that have an influence on any of the 6 diseases in our experiment (153 kinds of drugs). The complete heat map can be is in Additional file 3: Fig. S3. Then we perform validations on our prediction for each of the six disease conditions. Each of the six disease conditions has enough sample size which can make our validation result more confident. We extract a list of drugs from the drug indication information resources for each of the six disease conditions provided by SIDER. All of the drugs in the six lists are known to treat the six disease conditions respectively, so we assume them as the ground truth and further compare them with our prediction.

Evaluation of known drug-disease associations

The results of the prediction at K for 6 disease conditions are shown in Fig. 2. The figure demonstrates the precision of our prediction at $K \in \{5, 10, 15, 20\}$. For type 2 diabetes, stroke, heart attack and congestive heart failure, it is clear that most of the drugs can be mapped into the ground truth(SIDER drug list) when the K is small, the precisions of all the four disease conditions in the figure are greater than or equal to 0.8 when $K = 5$, their precision will decrease with the increase of K. However, the results of asthma and coronary heart disease were not as expected. For coronary heart disease, there is not so many known drug-disease pair in SIDER, which could be a reason for the low precision of this disease condition. Some of the clinical variables, like cholesterol, LDL (low-density lipoprotein), HDL (high-density lipoprotein) and triglycerides are more salient features than other clinical variables. So, our analysis for the disease condition which does not have a strong correlation with these clinical variables could have low precision. Apart from those known treatments of each target disease that can be found in the ground truth, there could be some unknown drug candidates which are likely to be the treatment of target disease.

The results of the FE test are shown in Fig. 3. As we can see, there is a negative linear relationship between the FE score and the group order. Since the average FE test score is decreasing with the ascending order of groups, so there is a positive linear relationship between FE score and average repurposing possibility score. The result in Fig. 3 shows that all 6 disease conditions demonstrate a negative linear relationship between their FE score and group order. Therefore, our scoring function is proven to be reasonable.

Case study and explainability

Having presented that our model successfully identified known associations between drugs and diseases, we further demonstrate the explainability of our model via corresponding complementarity between laboratory results of drug effect vectors and disease sign vectors. To exemplify this, we select 5 drug-disease association pairs (i.e.,

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 8 of 11

Type 2 diabetes-Clopidogrel hydrogen sulfate, Type 2 diabetes-Doxycycline hyclate, Coronary heart disease-Alendronate sodium, Congestive heart failure-Alendronate sodium and Heart attack-Alendronate sodium). For a given disease, the selected drugs are in its top-20 predicted list but not have been indicated as the treatment. In order to vividly compare the clinical vectors of the drug candidates and the corresponding disease, we present the clinical variables which contribute to their repurposing possibility score in Table 1. All the detailed clinical vectors can be found in Additional file 1: Table S2 and Additional file 2: Table S3. Combining the clinical disease vectors with clinical drug effect vectors, we can analyze why the drug candidates we select are potential treatments for corresponding disease conditions from the standpoint of clinical variables included in our experiment.

In the case of type 2 diabetes, we found that clopidogrel hydrogen sulfate could have a therapeutic effect on type 2 diabetes and Doxycycline Hyclate. Clopidogrel hydrogen sulfate is an antiplatelet medication and can be used to reduce the risk of myocardial infarction and stroke [21]. A study reported that clopidogrel will alleviate insulin resistance and improve glycemic control in type 2 diabetic patients [22], which is an important cause of insulin resistance. From the clinical drug effect vector of clopidogrel and clinical disease sign vector of type 2 diabetes, we can see clopidogrel and type 2 diabetes have the opposite effect on the cholesterol and LDL level. Lower cholesterol and LDL levels are biological markers of good glycaemic control [23], which is also corresponding to the literature study. Doxycycline Hyclate is an antibiotic which is primarily used to treat a wide range of bacterial infections. From the clinical vectors of Doxycycline and type 2 diabetes, we can see they have the opposite effect on the serum glucose level. High fasting blood glucose level is a common biological marker among type 2 diabetes patients. This finding is supported by a medical study that doxycycline can improve insulin resistance and fasting blood glucose level [24]. The analysis based on the opposite effect of type 2 diabetes and clopidogrel proves our prediction is reasonable, clopidogrel and doxycycline may be used as treatments for type 2 diabetes.

Alendronate sodium is usually used to treat osteoporosis [25]. We found it can potentially have a therapeutic effect on cardiovascular disease, including congestive heart failure, heart attack and coronary heart disease. Experiments show that alendronate can induce significantly lower cardiovascular mortality and reduce the risk of cardiovascular incidents [26]. A possible explanation given by this study is that bone and cardiovascular remodeling share some biological markers. From the clinical drug effect vectors of alendronate, we can see
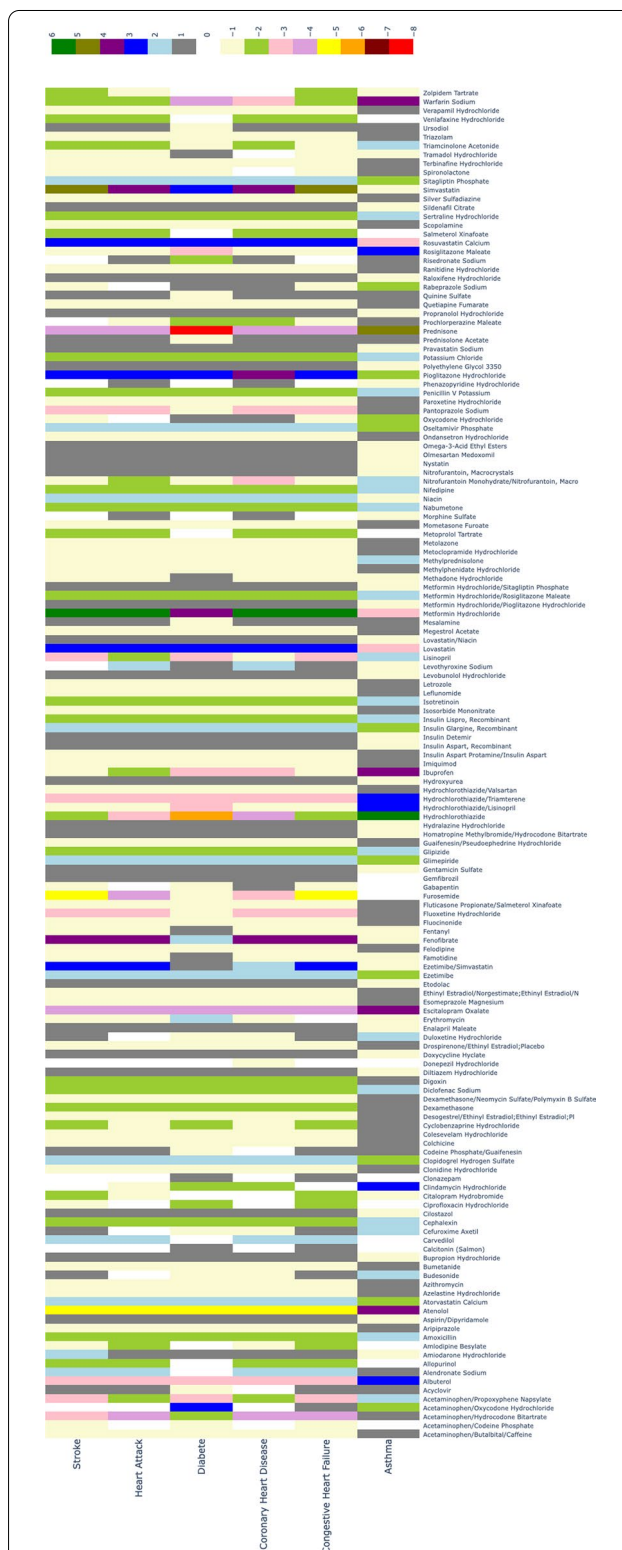

Fig. 4 Heat map of drug-disease repurposing possibility scores. X-axis stands for the 6 disease conditions and Y-axis is the name of the 153 drugs that have an influence on any of the diseases involved in our experiment. The color bar above the heat map annotates the scores that different colors in the heat map stand for

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 9 of 11

alendronate can lower alkaline phosphatase (ALP) and elevate the HDL level. Researches show that ALP can catalyze the inhibitor of vascular calcification, thus high-level ALP may lead to vascular hardening and promotes the atherosclerotic process [27]. On the other hand, HDL will promote reverse cholesterol transport, which could reduce the risk of cardiovascular events [28]. Thus, it seems possible that alendronate could be repurposed as a treatment for cardiovascular disease.

Highly related drugs and diseases

In Fig. 4, we demonstrate part of the repurposing possibility scores in the form of heat map. To further digging the relation within different drugs or diseases, we use bi-clustering algorithm to do a clustering for the drugs and diseases in Fig. 4. Bi-clustering is a data mining technology that simultaneous clustering of both row and column sets in a data matrix [29]. Given an $m \times n$, bi-clustering algorithm will generate new $m \times n$ matrix that a subset of rows which exhibit similar behavior across a subset of columns, or vice versa. In our work, we use bi-clustering algorithm to find different drugs with similar effect on some a disease and different diseases which can be treated with same kind of drug. The clustering result is plotted in Fig. 5. As shown in Fig. 5, type 2 diabetes has a strong correlation with heart diseases and stroke. We can also find many drugs that can decrease blood lipid or sugar level have a therapy effect on those diseases. In the further, these findings can help to find potential drug-disease pairs.

Limitations and further work

The verification results above show that our framework may identify some potential drug indications and thus help researchers find novel uses of existing drugs. However, our framework still has some limitations and space to improve.

First of all, we only include 6 kinds of diseases and 392 kinds of drugs in the our work. Actually, there are some other disease conditions and drugs that can be found in NHANES and EHR dataset. The reason we just include a part of drugs and diseases is that many of them have a small sample size so that we can not get a reliable result from them. To guarantee the results we get from the dataset are reliable enough, the sample size of each drug and disease that included in this work is larger than 1000. Due to this threshold, the experiments are conducted on 6 diseases conditions and 392 drugs, but the results we get are reliable and robust. In the future, we can include more drug-disease pairs with a large-scaled dataset.

The second limitation is the clinical variables involved in the experiment. Hundreds of clinical variables (laboratory results) can be found in the NHANES dataset, but
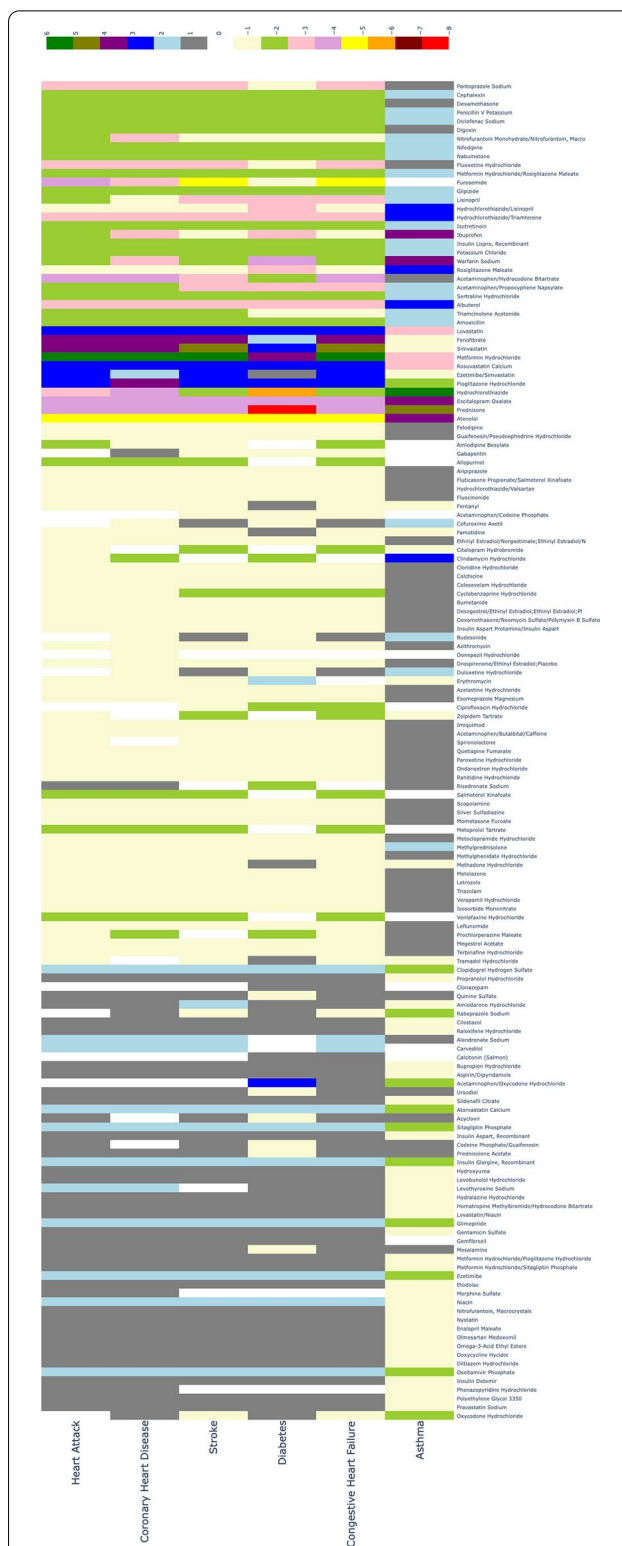

Fig. 5 The heat map after bi-clustering, X-axis stands for the 6 disease conditions and Y-axis is the name of the 153 drugs that have an influence on any of the diseases involved in our experiment. The color bar above the heat map annotates the scores that different colors in the heat map stand for

Wen *et al. BMC Med Inform Decis Mak* 2021, **21**(Suppl 8):263

Page 10 of 11

we still need to match them with the clinical variables in the EHR dataset. However, 35 kinds of clinical variables cannot completely reflect the human physiological activity, so it would be also addressed if we have a larger EHR dataset that contains more clinical variables.

Conclusion

In this paper, we establish a drug repurposing computational framework by using the electronic clinical information from the National Health and Nutrition Examination Survey (NHANES) and Electronic Health Records(EHR). We consider both of the opposite and same expressions between clinical disease sign vector and clinical drug effect vector in each drug-disease pair to calculate the repurposing possibility score. Our inferences of the novel use for different drugs are based on their repurposing possibility score with different disease conditions. We verify our predictions by fold-enrichment test and top K precision. Then, we further prove the feasibility of our model by doing a literature analysis of our prediction result. The result shows that our framework can not only retrieve the known indications of existing drugs but also find the previously unknown indications of existing drugs. So our framework can be potentially used in the drug repurposing tasks.

Abbreviations

NHANES: National Health and Nutrition Examination Survey; EHR: Electronic health record; LDL: Low-density lipoprotein; HDL: High-density lipoprotein; ALP: Alkaline phosphatase; CSCCS: Continuous self-controlled case series; FE: Fold enrichment.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-021-01617-4.

Additional file 1: Table S2. Clinical Disease Sign Vector. This table presents 6 kinds of disease conditions involved in our experiment and their influences on the 35 kinds of laboratory results

Additional file 2: Table S3. Clinical Drug Effect Vector. This table presents 392 kinds of existing drugs or drug combinations involved in our experiment and their influences on the 35 kinds of laboratory results

Additional file 3: Figure S3 Detailed Drug-Disease Heat Map. We transform the repurposing possibility score table into heat map and present it in this figure. This version includes the repurposing possibility scores of all the drug-disease pair.

Additional file 4: Table S4. Repurposing Possibility Score. This table include the repurposing possibility score of each drug-disease pair in our experiment

Additional file 5: Table S1. Laboratory Result List. This table includes the name of 35 kinds of laboratory result involved in ourexperiments and their corresponding NHANES code. **Figure S1.** Disease Clinical Variable Statistics. The figure present number of diseases will increase (Up) or decrease(Down) the level of each clinical variables. X-axis is the name of eachclinical variable, Y-axis is the number diseases. Blue bar stands for the "Up"relation, red bar stands for the "Down" relation. **Figure S2.**

Drug Clinical Variable Statistics. The number of drugs will increase (Up) or decrease (Down) the level ofeach clinical variables. X-axis is the name of each clinical variable, Y-axis isthe number diseases. Blue bar stands for the "Up" relation, red bar standsfor the "Down" relation.

Availability of data and materials

NHANES data analysed in the study is available on National Center for Health Statistics. The source code is provided for reproducing and is available at https://github.com/HoytWen/CCMDR.

Declarations

Ethics approval and consent to participate

As a secondary data analysis based on existing and deidentified data, this work was not classified as human subjects research and did not require Institutional Review Board approval.

Consent for publication

Not applicable.

Competing interests

PZ is the member of the editorial board of BMC Medical Informatics and Decision Making. The authors declare that they have no other competing interests.

Author details

[1]Department of Electrical and Computer Engineering, The Ohio State University, 2015 Neil Ave, Columbus, Ohio 43210, USA. [2]Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave, Columbus, Ohio 43210, USA. [3]Department of Biomedical Informatics, The Ohio State University, 1800 Cannon Drive, Columbus, Ohio 43210, USA.

References

1. O'Connor KA, Roth BL. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. Nat Rev Drug Discov. 2005;4(12):1005.
2. Chong CR, Sullivan DJ Jr. New uses for old drugs. Nature. 2007;448(7154):645.
3. DiMasi JA. New drug development in the united states from 1963 to 1999. Clin Pharmacol Ther. 2001;69(5):286–96.
4. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really 802 million? Health Aff. 2006;25(2):420–8.

5. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18(1):41–58.
6. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, et al. Predicting new molecular targets for known drugs. Nature. 2009;462(7270):175.
7. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35.
8. Lamb J. The connectivity map: a new tool for biomedical research. Nat Rev Cancer. 2007;7(1):54–60.
9. Luo H, Zhang P, Cao XH, Du D, Ye H, Huang H, Li C, Qin S, Wan C, Shi L, et al. Dpdr-cpi, a server that predicts drug positioning and drug repositioning via chemical-protein interactome. Sci Rep. 2016;6(1):1–9.
10. Wang Y, Chen S, Deng N, Wang Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLOS ONE. 2013;8(11):78518.
11. Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, Liu F. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinform. 2018;19(1):1–12.
12. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical RD. Nat Rev Drug Discov. 2011;10(6):428–38.
13. Jung J, Lee D. Inferring disease association using clinical factors in a combinatorial manner and their use in drug repositioning. Bioinformatics. 2013;29(16):2017–23.
14. Jang D, Lee S, Lee J, Kim K, Lee D. Inferring new drug indications using the complementarity between clinical disease signatures and drug effects. J Biomed Inform. 2016;59:248–57.
15. Kuang Z, Thomson J, Caldwell M, Peissig P, Stewart R, Page D. Computational drug repositioning using continuous self-controlled case series. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; p. 491–500. ACM.
16. Ghalwash M, Li Y, Zhang P, Hu J. Exploiting electronic health records to mine drug effects on laboratory test results. In: Proceedings of the 2017 ACM on conference on information and knowledge management, 2017; p. 1837–1846.
17. Cdc C. National health and nutrition examination survey. NCFHS (NCHS). US Department of Health and Human Services. Centers for Disease Control and Prevention. 2005.
18. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100(16):9440–5.
19. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol). 1996;58(1):267–88.
20. Kuhn M, Letunic I, Jensen LJ, Bork P. The sider database of drugs and side effects. Nucleic Acids Res. 2015;44(D1):1075–9.
21. Jarvis B, Simpson K. Clopidogrel. Drugs. 2000;60(2):347–77.
22. Taher MA, Nassir ES. Beneficial effects of clopidogrel on glycemic indices and oxidative stress in patients with type 2 diabetes. Saudi Pharm J. 2011;19(2):107–13.
23. Khan H, Sobki S, Khan S. Association between glycaemic control and serum lipids profile in type 2 diabetic patients: HbA 1c predicts dyslipidaemia. Clin Exp Med. 2007;7(1):24–9.
24. Wang N, Tian X, Chen Y, Tan H-Q, Xie P-J, Chen S-J, Fu Y-C, Chen Y-X, Xu W-C, Wei C.-j. Low dose doxycycline decreases systemic inflammation and improves glycemic control, lipid profiles, and islet morphology and function in db/db mice. Sci Rep. 2017;7(1):1–15.
25. Porras AG, Holland SD, Gertz BJ. Pharmacokinetics of alendronate. Clin Pharmacokinet. 1999;36(5):315–28.
26. Sing C-W, Wong AY, Kiel DP, Cheung EY, Lam JK, Cheung TT, Chan EW, Kung AW, Wong IC, Cheung C-L. Association of alendronate and risk of cardiovascular events in patients with hip fracture. J Bone Miner Res. 2018;33(8):1422–34.
27. Panh L, Ruidavets JB, Rousseau H, Petermann A, Bongard V, Bérard E, Taraszkiewicz D, Lairez O, Galinier M, Carrié D, et al. Association between serum alkaline phosphatase and coronary artery calcification in a sample of primary cardiovascular prevention patients. Atherosclerosis. 2017;260:81–6.
28. Rader DJ, Hovingh GK. HDL and cardiovascular disease. Lancet. 2014;384(9943):618–25.
29. Mirkin B. Mathematical Classification and Clustering, vol. 11. Berlin: Springer Science and Business Media; 2013.

Publisher's Note