Text Mining and Predicting Disease-Gene-Drug Associations of Hypertension Data Cubes-Based

Xing Wei
Bengbu Medical College
Xuelian Chang (xuelianchang@126.com)
Bengbu Medical College https://orcid.org/0000-0003-4974-2585
Yanqiu Wang
Affiliated Hospital of Bengbu Medical College
Jing Xie
Bengbu Medical College
Xiaodi Yang
Bengbu Medical College
Xiulin Jiang
Bengbu Medical College

Research

Keywords: Text mining, Hypertension, Data cube, Association extraction, Association Prediction

Posted Date: April 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-22783/v1

Abstract

Background Predispositions to hypertension is possibly associated with numerous potential gene polymorphisms and systemic disorders. Large-scale text mining of biomedical literature is a flexible and essential tool that can be applied to search for innovative drugs and treatments for diseases, such as investigating and predicting the bio-entities associations.

Result We proposed a generality approach for extracting and predicting hypertension-related disease-gene-drug associations based on dictionary and data cube from biomedical abstracts. After data preprocessing, we constructed the 0-D vertex cube, which we then filtered to construct three 1-D cubes consisting of 252 diseases, 185 genes, and 141 drugs. By applying association rules to quantify the disease-gene-drug associations, we found 235 associations between 79 diseases and the 71 genes, and AUCs was 84.1%; 196 associations between 43 diseases and 102 drugs, and AUCs was 85.8%; 160 associations between 31 genes and 106 drugs, and AUCs was 83.6%. Using the bottom-up computation algorithm, we established three 2-D cubes and one 3-D disease-gene-drug cube, which revealed 591 associations between 90 diseases, 82 genes, and 145 drugs. Based on this 3-D cube, we obtained 262 predictive bio-entity association pairs of which 57 disease-drugs, 84 disease-genes, and 121 gene-drugs.

Conclusions We have implemented and validated a data cube-based text mining approach to identifying and ranking the hypertension-related disease-gene-drug associations. Our results provide new pathways in the search for the potential treatment drugs of hypertension.

Introduction

Text mining has been established as a necessary NER tool help improve knowledge reusability from the large number of biomedical literature databases such as PubMed, which in turn creates new opportunities and challenges to explore the causes of diseases and their potential treatment drugs [,]. The rich biological/medical genomic databases provide a further opportunity to discover disease-gene-drug association. At present, automatically annotating biological entities such as diseases/genes and drugs and other important information in the biomedical literature is the first step in computational approaches to predicting associations between bio-entities, and it is useful for improving the scalability of biocuration services [,]. In order to identify semantic associations between biological/medical entities and extract structural associations from the rich literature databases, a dedicated set of computational and analytical techniques is required []. Some of the basic biological connections between entities, such as gene-gene associations [], drug-drug associations [], protein-drug association [] and drug-disease associations [], can be revealed through these computational and analytical approaches.

Analytical tools are essential to uncovering causal disease-gene-drug associations, and in this era of big data, data cubes and related analytical techniques can be useful tools. Data cubes can store multiple data dimensions (e.g., diseases, genes, and drugs) and measures (e.g., strength of association) in multidimensional ways [,]. Online analytical processing (OLAP), such as the following drill or roll-up

operation, performs multidimensional analysis of medical and genomic data and provides the capability for complex calculations, trend analysis, and sophisticated data modeling. Molecular, especially genomic, techniques have been widely used in biological and medical research [], producing mass amounts of results. These data provide an opportunity for exploring complex disease-gene-drug associations. In this context, data cubes and related association analytical methods can be good modeling and prediction tools.

Hypertension is an important worldwide public-health challenge because of its high frequency and concomitant risks of cardiovascular and kidney disease. []. Predispositions to essential hypertension is possibly associated with numerous gene polymorphisms [], for example, the gene ACE, an enzyme that plays a major role in blood pressure homeostasis and is an important target of anti- hypertensive drugs [], and several systemic disorders []. Drug selection is based on efficacy in lowering BP and in reducing cardiovascular end points including stroke, myocardial infarction, and heart failure []. Studies have also found that PDI as a potential therapeutic target in the treatment of atherosclerosis, thrombosis and hypertension []. While each of these mutations taken alone may cause only a slight impact, the combination causes severe changes in hypertension symptoms [,]. Identifying hypertension diseases-gene-drug associations will increase our understanding of the genetic pathogenesis and drug targets of hypertension, which will contribute to the development of novel prevention and treatment of hypertension in the future.

In this study, we probed the data in a multidimensional space based on a data cube structure and used association rules to determine the associations between bio-entities. We used hypertensive as a model disease for the construction of disease-gene-drug data cube to analyze the potential associations between diseases, genes, and drugs. The results provide valuable information for the development of innovative diagnosis and treatment tools for hypertension based on candidate genes.

Materials And Methods

Dictionary construction

Dictionary-based methods instead rely – as the name suggests – on matching a diction of names against text. For this purpose, the quality of dictionary is obviously very important, the best performing methods for NER according to blind assessments rely on carefully curated dictionaries to eliminate synonyms that give rise to many false positives []. Moreover, dictionary-based methods have the crucial advantage of being able to normalize names. A high-quality, comprehensive dictionary of disease / gene / drug names is thus a prerequisite for mining disease-gene-drug associations from the bio-literature.

We used the ICD10 [], Entrez GENE [], Gene Ontology [], OMIM [] and DrugBank [] databases to compile standardized gene and drug dictionaries, which we named “DiseaseDictionary” (with entries for 26,813 human diseases) “GeneDictionary” (with entries for 40,172 human genes) and “DrugDictionary” (with entries for 1,763 drugs), respectively. Each entry included the disease’s /gene’s / drug’s standard name,

alias, synonyms, standard code, etc. ICD 10 is designed for clinical coding and billing purposes; its structure and disease names are poorly suited for bio-literature mining.

To improve recall, we automatically generated variants of the disease names. Although the terms *disease, disorder*, and syndrome have separate definitions, we found that they are used inconsistently in the literature when part of disease names; for instance, *Alzheimer's disease* is occasionally referred to as *Alzheimer's disorder* or *Alzheimer's syndrome*. We also removed words in parentheses and brackets occurring at the end of disease names, unless this would cause ambiguity.

Integration of Corpus

We searched the diabetes-related literature in PubMed for the most recent year using the following search strategy: "((hypertension) AND (("1988/3/16"[Date - Publication]: "2018/3/16"[Date - Publication])))". The search returned a total of 334,155 with abstracts, which we downloaded and saved in text format.

On account of the literature abstract unstructured data, we needed to standardize it before proceeding with our analysis. We used the normalization method, one of the most widely used approaches for the extension of dictionaries with synonyms []. In this study, all entity sets were standardized based on the names of genes (drugs) extracted from the literature.

Corpus Preprocessing

Here, we describe how to preprocess the corpus by applying three text processing steps: sentence splitting, tokenization and recognition. This step is the text corpus which is used for the task of association classification and the output of this step is the set of unigram features that will be further used for the feature set. The preprocessing activities used in this process are:

1)
Sentence splitting: the sentence splitter splits the text that is required for taggers into sentences. The sentence splitter uses a dictionary list of abbreviations to differentiate between full stops and other token types. Sentence splitter takes the "(.)" to split one sentence from another. For example, "prevention, diagnosis and treatment" is a single sentence. 2,864,308 sentences were obtained.
2)
Tokenization: the tokenizer splits the text into small tokens, that was, different type of words, punctuation, and numbers. For instance, "prevention, diagnosis and treatment" had eight tokens, that was, (prevention), (,), (space), (diagnosis), (space), (and), (space), and (treatment).
3)
Recognition: to match a document against the dictionary, we have labeled each word with the method of BIO (B-begin, I-inside, O-outside, E-end) [], so that it became a standard text corpus for entity recognition and association extraction. For instance, for the statement "However, in patients with *type 2 diabetes* and *hypertension*, multiple studies demonstrate the benefit of gene ACEI or ARB in preventing or delaying the onset of nephropathy.", the classification results are shown in Fig. 1. For each name recognition in the

text, we normalize it to the corresponding unique identifier and, in case of diseases, backtrack the term to the root of the ontology through “is_a” relationship to assign also the identifiers of all parent terms.

Extraction and sorting of bio-entities association

After corpus preprocessing, we extracted and scored associations between disease, gene and drug using association rules of “*Support*”, “*Confidence*” and “*Lift*” , which take into account co-occurrences at the level of individual sentences, simultaneously. For bio-entity A and B (such as: disease, gene, and drug), “*Support*” measures the frequency of a bio-entity against the total corpus:

$$Support(A \Longrightarrow \mathrm{B}) = \mathrm{P}(\mathrm{A} \cup \mathrm{B}) = a/N$$

where *a* is the number of occurrences and *N* is the total number of sentences / associations in the corpus / network.

“*Confidence*” measures the intensity of the association:

$$Confidence(\mathrm{A} \Longrightarrow \mathrm{B}) = \mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{P}(\mathrm{A} \cup \mathrm{B})}{P(A)}$$

“*Lift*” assesses whether a forecasting model is effective and reflects the importance of set to set :

$$Lift(\mathrm{A} \Longrightarrow \mathrm{B}) = \frac{Confidence(\mathrm{A} \Longrightarrow \mathrm{B})}{P(A)} = \frac{\mathrm{P}(\mathrm{A} \cup \mathrm{B})}{\mathrm{P}(\mathrm{A})\mathrm{P}(\mathrm{B})}$$

If the value of *Lift* is 1, A and B are not associated; if *Lift* < 1, the emergence probability of A is inversely proportional to B; if *Lift* > 1, the higher the value, the stronger the association between A and B []. Considering that related entities might be mentioned only occasionally or in comparison to one another in the abstract, we set the minimum threshold of *Lift* to 3. This is equivalent to a confidence level above 99.8% or a critical value of 3 times the standard deviation in a standard normal distribution; i.e., *Lift* > 3 is considered a strong association.

Data cube

Data cubes are defined by facts (the data elements being measured) and dimensions (the perspectives from which data is analyzed). Each dimension has an associated table, known as a dimension table. In this study, we used the biomedical documents downloaded from PubMed as a data warehouse, in which the biomedical entities (disease, gene, and drug) were the dimensions. To explore the associations between different dimensions, we used the values of support and lift as measures of fact.

A total of eight cubes (or groups) comprise all possible combinations of disease, gene, and drug, including the empty set. For instance, (disease, gene, drug) represents the disease-gene-drug cube, and (disease, gene) represents the disease-gene cube (Fig. 2). The vertex cube (or 0-D cube), often described as “all,” is the most generalized (least specific), while the basic cube is the least generalized (most specific).

Bottom-up computation (BUC) algorithm

The diabetes data cube in this study is an iceberg cube, so it is suitable for using the BUC algorithm to build the network model of associations. The details of the BUC algorithm have been described elsewhere []. Briefly, the algorithm drills down from the top, i.e., from higher-level, less detailed units to lower-level, more detailed units. In this study, we used *Lift* as the measure of association for partitioning. Along the recursive process, frequent combinations are sent to output. When all attribute values are partitioned in the last dimension, the algorithm recurses back to the previous level, and the next attribute value is processed, and so on. The algorithm eventually returns a full association network.

ROC curve

The receiver operating characteristic (ROC) curve, which detects the accuracy of a binary classification algorithm, is widely used to evaluate the performance of medical diagnostic tests []. The accuracy of a test is measured by the area under the ROC curve. In this study, we used R v3.3 to construct the network, create the ROC curve, and evaluate the algorithm performance. R is a computer language and environment for statistical computing and graphics. It provides more than 5,000 open source packages (e.g., igraph) that enable the construction and analysis of networks, as well as a universal tool offering network operation data and network algorithm implementation.

Results And Discussion

0-D Vertex cube

After data preprocessing, the standardized all-entity set is the 0-D vertex cube of the hypertension data cube (252 Disease corpus, 159 Gene corpus & 141 Drug corpus); this is the data foundation for further research in this study.

1-D Disease, gene, and drug cube

Due to the very low frequency of some genes and drugs in the literature, we excluded genes and drugs with frequency < 0.1%. We filtered the 0-D cube against DiseaseDictionary, GeneDictionary and DrugDictionary. The resultant 1-D gene cube consisted of 252 diseases, 1-D gene cube consisted of 185 genes and the 1-D drug cube consisted of 141 drugs (Table 1, Attachment 1-D.xlsx).

Table 1

Partial list of 1-D disease, gene and drug cube of hypertension

No.	Disease	Fre.	Gene	Fre.	Drug	Fre.
1	Atherosclerosis	11645	ACE	8636	Glucose	16362
2	Coronary Artery Disease	8863	TNF	1428	Insulin	15259
3	Cardiovascular Diseases	6318	PTH	698	Calcium carbonate	14786
4	Atrial Fibrillation	4546	AGT	595	Nitric Oxide	10499
5	Cardiomyopathy	3289	HBA1	574	Oxygen	9046
6	Aneurysm	3194	EGFR	521	Triglycerides	5590
7	Apnea	3189	ACE2	349	Norepinephrine	5134
8	Anemia	2979	RHOA	338	Captopril	3888
9	Arthritis	2219	BMPR2	303	Hydrochlorothiazide	3283
10	Ascites	2148	MTHFR	291	Nifedipine	3185
11	Asthma	2045	MTOR	184	Uric Acid	2865
12	Bradycardia	1917	NOX4	182	Propranolol	2860
13	Anxiety	1818	NOS3	181	Losartan	2788
14	Arrhythmia	1817	IRS	173	Enalapril	2593
15	Angina Pectoris	1714	WNK4	164	Amlodipine	2361
16	Coronary Disease	1363	WNK1	150	Nitroprusside	2321
17	Arteriosclerosis	1152	CTGF	148	Atenolol	2274
18	Coma	1100	STAT3	139	Phenylephrine	2242
19	Cough	1087	ADM	127	Dopamine	1981
20	Acidosis	1028	ALB	127	Epinephrine	1742

The top three diseases of the 1-D cube are *Atherosclerosis*, *Coronary artery disease* and *Cardiovascular disease*, which are closely related to *Hypertension*. For instance, *Atherosclerosis* is an inevitable result of degenerative aging, is the main cause of *heart disease* and *stroke* [], and often coexists with *hypertension* [].

No.	Disease	Fre.	Gene	Fre.	Drug	Fre.
There were few studies on the associations between hypertension and gene, and ACE was the most important one enzyme [], which was involved in the process of transforming ACE I into ACE II with physiological activity, may be related to myocardial infarction, SARS resistance, renal diabetes, Alzheimer's disease and other diseases. TNF gene was expressed in leukocytes and macrophages; it was involved in protein nuclear entry, positive regulation of protein amino acid phosphorylation, negative regulation of L-glutamate transport, glucose metabolism, etc.; it was involved in dementia, migraine, asthma susceptibility, septicemia susceptibility and other diseases.						
Glucose was the most frequently studied drug related to hypertension. Its "access number" in DrugBank was "db09341". It was mainly stored in animals as plant starch and glycogen. It helped various metabolic processes at the cell level, usually in the form of injection, providing nutritional supplement for metabolic disorder or improper regulation of blood glucose level. *Glucose* is one of the most important drugs in the World Health Organization (WHO) list of essential drugs.						

2-D Disease-gene cube

Table 1 provides the *support* and *lift* score between disease and gene in the 2-D cube.

Table 1

Top 10 of disease-gene associations and sorting by *lift*.

Rank	Disease	Gene	Co-occurrence	*Support*	*Lift*
1	Hypertension	ACE	71	0.302	114.8
2	CADASIL	NOTCH3	28	0.119	82.1
3	Angioedema	ACE	26	0.111	78.5
4	Cough	ACE	21	0.089	73.0
5	Hypertension	EGFR	21	0.089	69.2
6	Hypertension	AGT	13	0.055	64.3
7	Proteinuria	EGFR	11	0.047	58.7
8	Stroke	ACE	10	0.043	50.1
9	Proteinuria	ACE	8	0.034	48.6
10	Stroke	EGFR	8	0.034	43.0

Using the association rules, 235 significant associations between the 79 classes of hypertension-related clinical manifestations / symptoms and the 71 candidate genes are extracted from the 2-D cube (see Fig. 3 (a), attachment 2-D.xlsx). We found that *Hypertension*, ACE and EGFR were belong to the central nodes of the 2-D cube and were associated with 21 genes, 35 and 19 diseases respectively. There were four pairs nodes of *one-against-one* strategy [], including: *Amenorrhea*-CYP17A1, Goiter-LMNA, *Melanoma*-BRAF and (*Pre-Eclampsia*)-FLT1-*Eclampsia*.

Using the *Support* method, we have grouped and sorted diseases / gene respectively in the 2-D cube. Top three of the sorting disease nodes are *Hypertension* (*Support = 0.213*), *Atherosclerosis* (*Support = 0.051*) and *Inflammation* (*Support = 0.047*). During algorithm development, we found that *Atherosclerosis, Inflammation, Obesity* and *Smoking* were associated with 12 genes (such as EGFR, TNF and CORIN, etc.), 10 genes (such as IL6, CCL2 and TNF, etc.), 10 genes (such as TNF, IRS and ACE, etc.) and 10 genes (such as ApoA1, ACE and ARMS2, etc.) respectively. Similarly, ACE (*Support = 0.157*), EGFR (*Support = 0.081*) and NTF (*Support = 0.072*) are the top three genes nodes. Among them, EGFR and NTF gene are associated with 19 diseases (such as *Anemia, Atherosclerosis* and *Depression*, etc.) and 15 diseases (such as *Apnea, Atherosclerosis* and *Castration*, etc.) respectively.

2-D Disease-drug cube

Table 2 provides the *support* and *lift* score between disease and drug in the 2-D cube.

Table 2

Top 10 of disease-drug associations and sorting by lift.

Rank	Disease	Drug	Co-occurrence	*Support*	*Lift*
1	Hypertension	Glucose	130	0.032	109.7
2	Obesity	Glucose	78	0.019	79.4
3	Hypertension	Losartan	71	0.017	73.0
4	Hypertension	Water	55	0.014	63.5
5	Hypertension	Nitric Oxide	52	0.013	59.8
6	Hypertension	Amlodipine	51	0.013	57.1
7	Hypertension	Potassium	49	0.012	55.5
8	Hypertension	Nifedipine	40	0.010	49.3
9	Stroke	Perindopril	39	0.010	47.1
10	Stroke	Warfarin	38	0.009	45.4

196 significant associations between 43 diseases and 101 drugs in the 2-D disease-drug cube (see Fig. 3 (b), attachment 2-D.xlsx).We found that the central nodes in this 2-D cube were the disease nodes of *Hypertension* (*Support = 0.383*), *stroke* (*Support = 0.122*), and the drug node of *Glucose* (*Support = 0.066*), which were associated with 75, 24 and 14 different types of entities respectively.

There are three pairs of *one-against-one* strategy between the disease and drug, including: *Pain NOS-Morphine*, *Celecoxib-AIDS-Diclofenac* and *Dorzolamide-Glaucoma-Timolol*.

2-D Gene-drug cube

Table 3 provides the *support* and *lift* score between gene and drug in the 2-D cube.

Table 3

Top 10 of gene-drug associations and sorting by lift.

Rank	Gene	Drug	Co-occurrence	*Support*	*Lift*
1	ACE	Captopril	715	0.054	61.0
2	ACE	Enalapril	708	0.053	49.2
3	ACE	Perindopril	381	0.029	47.8
4	ACE	Ramipril	381	0.029	42.2
5	ACE	Lisinopril	339	0.026	39.1
6	ACE	Losartan	305	0.023	37.2
7	ACE	Glucose	222	0.017	25.7
8	ADM	Nitric Oxide	216	0.016	25.5
9	ACE	Quinapril	201	0.015	23.7
10	ACE	Amlodipine	185	0.014	20.4
160 significant associations between 31 genes and 104 drugs in the 2-D disease-drug cube (see Fig. 3 (c), attachment 2-D.xlsx), and the ACE (*Support = 0.519*) gene is the central node, which is associated with 83 drugs. TNF (*Support = 0.087*) and *Glucose* (*Support = 0.081*) are associated with 14 drugs and 13 genes secondly. There are 15 gene nodes with unique association, such as AGT, FGF23, GRK4, and so on. In addition, 83 drug nodes with unique association, 75 of which are related to ACE gene.					
In this cube, there are five gene-drug pairs of *one-against-one* strategy, including: MMP2-*Doxycycline*, GRK4-*Dopamine*, HFE-*Iron*-FGF23, DICER1-*Progesterone* and MYD88-*Colchicine*; *Nitric Oxide* is associated with S100B and NOS2 gene; NGF is associated with 9 drugs; CORT is associated with 7 drugs; The APP and CYP2C8 gene are associated with 4 drugs. *Procaine* can inhibit the expression of STAT3 at mRNA and protein levels. It is a potential therapeutic drug for the treatment of neuropathic pain [].					

3-D Disease-gene-drug cube

Based on the 1-D and 2-D cubes and their associated intensities and networks, we constructed the disease-gene-drug network of hypertension. We used association rules and the *Lift* threshold to estimate whether there was a significant association between diseases, genes and drugs, and we calculated the association strength. After removing data duplications, we found 591 associations between 90 diseases, 82 genes, and 145 drugs. The 3-D disease-gene-drug network of hypertension is shown in Fig. 4.

We found *Hypertension*, ACE and *Stroke* are the central nodes in the 3-D cube, which are associated with 123, 118 and 32 other nodes respectively (attachment 3-D.xlsx).

Using the *Support* method, we have grouped and sorted diseases / gene / drug respectively in the 3-D cube. Top three of the sorting disease nodes are *Hypertension* (*Support = 0.208*), *Stroke* (*Support = 0.054*)

and *Atherosclerosis* (*Support = 0.037*), which are associated with 123, 32 and 22 nodes respectively; 41 nodes are association with unique node, including *Bradycardia*, *Coronary Disease*, *Dehydration*, *Diabetic Retinopathy*, *Glucose Intolerance*, etc.

The top three nodes of the sorting drug nodes are *Glucose* (*Support = 0.044*), *Oxygen* (*Support = 0.022*), *Nitric Oxide* (*Support = 0.015*) and *Losartan* (*Support = 0.015*), which are associated with 26, 13, 9 and 9 nodes respectively; 72 nodes are association with unique node, including: *L-Argine, Lacidipine, Camphor, Bezafibrate* and *Bezafibrate*, etc.

The top three nodes of the sorting gene nodes are ACE (*Support = 0.200*), TNF (*Support = 0.049*) and EGFR (*Support = 0.041*) are the top three nodes in the 3-D cube, which are associated with 118, 29 and 24 nodes respectively. There are 38 nodes are association with unique node, including: CYP4A11, HMOX1, BRAF, CCR2, RGS2, etc.

There are five pairs associations of *one-against-one* strategy besides the network model, including: *Goiter* - LMNA, *Amenorrhea* - CYP17A1 and *Melanoma*-BRAF, (*Pre-Eclampsia*) - FLT1 - *Eclampsia* and *Glaucoma* – *Dorzolamide* - *Timolol*.

Evaluation of association using ROC curve

We used ROC curve analysis to evaluate the ability of hypertension-related disease-gene-drug association to discriminate true positive (TP) from false positive (FP) hidden associations using literature partitioning (see Fig. 5). The area under the curve (AUC) was used as the accuracy indicator. Associations were validated using the following criteria: 1) TP = direct association and co-publications at least 3, e.g., *Hypertension* and ACE; and 2) FP = no direct association or co-publications < 3, e.g., *Coma* and ADM. The AUCs were (mean ± standard error) 0.842 ± 0.032, 0.858 ± 0.044, and 0.836 ± 0.045, and the asymptotic 95% confidence intervals were (0.778, 0.903), (0.773, 0.944), and (0.748, 0.924) for disease-gene, disease-drug, and gene-drug, respectively.

Taken together, these results show that our methods are robust and can be applied to quickly detect a variety of biologically relevant hidden associations. As with other studies of association extraction algorithms [], we also obtained some predictions as results. This is also a goal of biomedical entity association extraction: to propose hypotheses and assist researchers in designing the direction of related experiments [].

Prediction of disease-gene-drug associations

In this study, we used ABC discovery method [] to predict hypertension candidate diseases, genes and drugs, and to mine new association between hypertension related diseases, genes and drugs. Similarly, some unproven bio-entity association pairs are also obtained in this study, and the predictive results with false-positive are allowed [], because this is also one of the objectives of association mining: to put forward predictive research hypotheses, to help biological researchers develop novel ideas, so as to design innovative experimental directions [].

This study verifies all the associations between disease-gene, disease-drug, and gene-drug. After verification, 262 kinds of predictive association (i.e. false-positive association) were obtained, including 57 kinds of disease-drug, 84 kinds of disease-gene and 121 kinds of gene-drug. Table 4 provides a partial list of disease-gene-drug associations linked implicitly but not explicitly to hypertension through the literature.

Table 4
Partial list of predictive associations between bio-entities.

Association	Entity1	Entity2
Disease-gene	Depression	SLC2A9
	Coma	SLC2A9
	Coma	AGT
	Pheochromocytoma	GRK4
Disease-drug	Anemia	Chlorthalidone
	Dementia	Benazepril
	Depression	Eplerenone
	Glomerulonephritis	Doxazosin
	Obesity	Spirapril
Gene-drug	ABCB1	Candesartan
	ABCB1	Clonidine
	APOA1	Amlodipine
	EGFR	Telmisartan
	MTHFR	Eprosartan
In the prediction of novel disease-gene association, for example, no study has yet reported whether an association exists between *Depression* and SLC2A9. The rs6855911 allele variation in SLC2A9 gene expression is strongly associated with serum uric acid concentration [], while the serum uric acid level in human body has a positive correlation with depression and anxiety; therefore, there may be potential association between the two bio-entity. AGT gene is expressed in adipose tissue, adrenal gland, brain, blood vessel and nervous system, which is involved in cell growth, positive regulation of cytokine synthesis, apoptosis and cardiac hypertrophy. AGT is a susceptible gene of hypertension, while hypertensive cerebral hemorrhage will cause coma; AGT is related to insulin resistance [], similarly, diabetes may also cause coma. It is suggested that AGT may be related to *Coma*.		

In the prediction of novel disease-drug association, there was a clinical case report [] that a 58 year old woman who was hospitalized with intermittent fever, accompanied by anemia and inflammation, had been taking atenolol and chlorothiazide to treat hypertension, and had increased diuretics due to uncontrolled hypertension six weeks before hospitalization; however, after discontinuing antihypertensive

drugs, the fever symptoms relieved rapidly, and the diagnosis might be allergic to chlorothiazide. Therefore, *Chlorothiazide* may be associated with *Anemia* and may cause fever. For the related bio-entity pair of *Glomerulonephritis-Doxazosin*, three cases report [] of primary hypertension in children caused by renal diseases, including renal atrophy, hydronephrosis secondary to reflux nephropathy, nephrotic syndrome, and acute streptococcal angio coccal nephritis, were reported in the literature. After taking doxazosin and other drugs to control hypertension, the patients recovered to health. Therefore, there may be some associations between the *Glomerulonephritis* and *Doxazosin*.

In the prediction of novel gene-drug association, ABCB1 is ATP binding family B1. At present, the mechanism of action between ABCB1 and *Candesartan* / *Clonidine* is not clear. ATP binding protein transporters, such as P-glycoprotein (P-gp / ABCB1) and multidrug resistance associated protein (MRP / ABCCs), are involved in the regulation of drug absorption, distribution and excretion. Candesartan is used to treat essential hypertension; clonidine is a kind of psychoactive drug, which can be used to treat severe hypertension; they can be combined with other drugs. However, the interaction between them and their effect on ABC transporters are still unclear. Some studies have shown that candesartan cilexetil can significantly inhibit P-glycoprotein activity []. Therefore, ABCB1 may be the candidate gene of *Candesartan* / *Clonidine*.

Conclusions

In this study, we developed a dictionary and data cube-based NER tool to extract and predict hypertension-related disease-gene- drug associations from PubMed. The results of our study provide novel predictions of disease-gene-drug associations, which will aid researchers in designing future medical/biological experiments.

Conceptual modeling of spatial data cubes requires that two types of metadata be defined: 1) metadata describing a data warehouse that consists of various data sources, can be maintained and integrated, and has model data structure; and 2) metadata describing how the data in the warehouse can be analyzed to meet the needs of decision makers. In the data cube of Hypertension, the entities we defined (diseases, genes, and drugs) can be viewed as the first type of metadata, while other entities as defined in literature and dictionaries can be viewed as the second type of metadata. Therefore, the model of the multidimensional dataset is network-based.

Using hypertension-related literature abstracts from the most recent year based on the entities of diseases, genes, and drugs dictionaries, we applied data cube analysis and association rules to highlight and extract the key nodes in the disease-gene-drug cubes network. The AUCs of our algorithm achieved 0.841, 0.858, and 0.836, respectively. This allowed us to mine the potential links between bio-entities, and make quantitative assessments. Meanwhile, the whole heterogeneous network may contribute further to the discovery of candidate genes and drugs in hypertension-related diseases and the deduction of novel disease-gene-drug associations. The next step will be to assess the performance of the algorithm on massive data sets and to promote its further use.

Abbreviations

ACE: angiotensin I converting enzyme (peptidyl-dipeptidase);

ADM: Adrenomedullin;

AUC: Area under the curve;

NER: Named entity recognition;

Pain NOS: Pain no other specified;

ROC: Receiver operating characteristic.

Declarations

Acknowledgements

We would like to appreciate anonymous reviewers for their valuable comments on the manuscript.

Funding

This study was funded by the Key Projects of Science Research in Education Institutes of Anhui Colleges and Universities (KJ2019A0320、KJ2019A0325、KJ2019A0331), Anhui University Excellent Youth Talent Support Program Project(gxyq2017033), the Overseas Outstanding Young Talents in Colleges and Universities(gxgwfx2019031), the Key Natural Science Foundation of Bengbu College (BYKY1825ZD)

Availability of data and materials

The disease dataset used in this article is available in the International Classification of Diseases (ICD 10) at https://icd.who.int/browse10/2019/en.

The gene dataset used in this article is available in the Gene Expression Omnibus (GEO), accession number GSE15852 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15852).

The drug dataset used in this article is available in the Drug Bank, accession number Accession Number DB05207 (https://www.drugbank.ca/releases/latest)

Authors' contributions

Xing Wei conceived and designed the study. Xuelian Chang analyzed and wrote the paper, and was a major contributor in writing the manuscript. Yanqiu Wang and Xiaodi Yang reviewed and edited the manuscript. Jing Xie and Xiulin Jiang provided the data. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1] Anhui Key Laboratory of Digital Medicine and Intelligent Health, Bengbu Medical College, Bengbu, Anhui 233030, China. [2] Anhui Key Laboratory of Infection and Immunity, Bengbu Medical College, Bengbu, Anhui 233030, China. [3] Department of Endocrinology, Second Affiliated Hospital of Bengbu Medical College, Bengbu, Anhui 233030, China.

References

1. Bourne PE, Lorsch JR, Green ED. Perspective: Sustaining the big-data ecosystem[J]. Nature. 2015; 527(7576): 16-17.
2. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics. 2012; 13(8): 523-36.
3. Alvin R, Jeffrey D, Isaac K. Machine Learning in Medicine [J]. The New England Journal of Medicine. 2019; 14(380): 1347-1358.
4. Islamaj R, Wilbur WJ, Xie N, et al. PubMed Text Similarity Model and its application to curation efforts in the Conserved Domain Database. Database (Oxford). 2019: baz064. doi:10.1093/database/baz064
5. David W, Hans-Henrik S, Christian T, et al. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts[J]. PLOS Computational Biology. 2018; 14(2): e1005962.
6. Lee S, Son D, Kim Y. et al. Unified Cox model based multifactor dimensionality reduction method for gene-gene interaction analysis of the survival phenotype. BioData Mining 2018; 11: 27.
7. Bui QC, Sloot PM, van Mulligen EM, et al. A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics. 2014; 30(23):3365-3371.
8. Chapy H, Goracci L, Vayer P, et al. Pharmacophore-based discovery of inhibitors of a novel drug/proton antiporter in human brain endothelial hCMEC/D3 cell line[J]. British Journal of Pharmacology, 2015; 172(20): 4888-4904.

9. Xu R, Wang QQ. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinformatics. 2013; 14(13):1-11.
10. Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. Nucleic Acids Research. 2015; 44(Database issue): D1036-D1044.
11. Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research. 2016; 45(Database issue): D566-D573.
12. Liu X, Cheng J, Zhang G, et al. Engineering Yeast for the Production of Breviscapine by Genomic Analysis and Synthetic Biology Approaches[J]. Nature Communications. 2018; 9(1): 448.
13. Patricia M Kearney, Megan Whelton, Kristi Reynolds, et al. Global Burden of Hypertension: Analysis of Worldwide Data[J]. Lancet. 2005; 365(9455):217-223.
14. Tsukada K, Ishimitsu T, Teranishi M, et al. Positive association of CYP11B2 gene polymorphism with genetic predisposition to essential hypertension[J]. Journal of Human Hypertension. 2002; 16(11): 789-793.
15. Jacob HJ, Lindpaintner K, Lincoln SE, et al. Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat[J]. Cell. 1991; 67(1): 213-224.
16. Tsukada K, Ishimitsu T, Teranishi M, et al. Positive association of CYP11B2 gene polymorphism with genetic predisposition to essential hypertension[J]. Journal of Human Hypertension. 2002; 16(11): 789-793.
17. Gradman AH, Basile JN, Carter BL, et al. Combination therapy in hypertension [J]. Journal of the American Society of Hypertension. 2010; 4(1):42-50.
18. Lopes LR, Trevelin SC. Protein disulfide isomerase and Nox: new partners in redox signaling[J]. Current Pharmaceutical Design. 2015; 21(41): 5951-5963.
19. Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension[J]. Annals of Medicine. 2002; 34(2): 88-95.
20. Qian X, Guo D, Zhou H, et al. Interactions Between PPARG and AGTR1 Gene Polymorphisms on the Risk of Hypertension in Chinese Han Population.[J]. genetic testing & molecular biomarkers. 2018; 22(2): 90-97.
21. Gaudan S, Kirsch H, Rebholz-Schuhmann D. Resolving abbreviations to their senses in Medline[J]. Bioinformatics. 2005; 21(18): 3658-3664.
22. Lash TL, Christensen S, Christiansen CF, et al. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients[J]. BMC Medical Research Methodology. 2011; 11(1): 83-89.
23. Donna M, Jim O, Pruitt KD, et al. Entrez Gene: gene-centered information at NCBI[J]. Nucleic Acids Research. 2007; 39(Database issue): D54-D58.
24. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology[J]. Nature Genetics. 2000; 25(1): 25-29.

25. Ada H, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders[J]. Nucleic Acids Research. 2005; 33(Database issue): D514-D517.
26. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. Nucleic Acids Research. 2017; 46(Database issue): D1074-D1082.
27. Khalid MA, Jijkoun V, Rijke MD. The Impact of Named Entity Normalization on Information Retrieval for Question Answering. Lecture Notes in Computer Science. 2008; 4956(4):705-710.
28. Barbara S. Between Species: Science and Subjectivity[J]. Configurations, 2008, 14(1-2):115-126.
29. Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data[J]. Knowledge and Information Systems. 2006; 9(3):1-2.
30. Mcnicholas PD, Murphy TB, Regan M O'. Standardising the lift of an association rule[J]. Computational Statistics & Data Analysis. 2008; 52(10):4712-4721.
31. Beyer K, Ramakrishnan R. Bottom-up computation of sparse and Iceberg CUBE[J]. Acm Sigmod Record. 1999; 28(2):359-370.
32. Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29-36.
33. Lusis AJ. Atherosclerosis[J]. Nature. 2000; 407(6801): 233-241.
34. Vincenti F, Amend WJ, Abele J, et al. The role of hypertension in hemodialysis-associated atherosclerosis[J]. American Journal of Medicine. 1980; 68(3):363-369.
35. Rigat B, Hubert C, Alhenc-Gelas F, et al. An insertion/Deletion polymorphism in the angiotensin I-Converting enzyme gene accounting for half the variance of serum enzyme level[J]. Journal of Clinical Investigation. 1990; 86(4): 1343-1346.
36. Hsu CW, Lin CJ. A Comparison of Methods for Multiclass Support Vector Machines[J]. IEEE Transactions on Neural Networks. 2002; 13(2):415-425.
37. Li D, Yan Y, Yu L, et al. Procaine Attenuates Pain Behaviors of Neuropathic Pain Model Rats Possibly via Inhibiting JAK2/STAT3[J]. Biomolecules & Therapeutics. 2016; 24(5): 489-494.
38. Sun K, Liu H, Yeganova L, et al. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. Journal of Biomedical Informatics. 2015; 55: 23-30.
39. Gonzalez GH, Tahsin T, Goodale BC, et al. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. Briefings in Bioinformatics. 2016; 17(1): 33-42.
40. Frijters R, Vugt MV, Smeets R, et al. Literature mining for the discovery of hidden connections between drugs, genes and diseases[J]. PLOS Computational Biology. 2010; 6(9): 655-664.
41. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain[J]. Methods. 2015; 74(3): 97-106.
42. Gonzalez GH, Tahsin T, Goodale BC, et al. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery[J]. Briefings in Bioinformatics. 2016; 17(1): 33-42.

43. Lyngdoh T, Bochud M, Glaus J, et al. Associations of Serum Uric Acid and SLC2A9 Variant with Depressive and Anxiety Disorders: A Population-Based Study[J]. PLOS One. 2013; 8(10): e76336.
44. Roubícek T, Dolinková M, Bláha J, et al. Increased Angiotensinogen Production in Epicardial Adipose Tissue during Cardiac Surgery: Possible Role in a Postoperative Insulin Resistance[J]. Physiological Research. 2008; 57(6): 911-917.
45. Osterwalder P, Koch J, Wüthrich B, et al. Unklarer Status febrilis[J]. Dtsch Med Wochenschr. 1998; 123(24): 761-765.
46. Akın F, Kılıçaslan C, Solak ES, et al. Posterior reversible encephalopathy syndrome in children: report of three cases[J]. Childs Nervous System Chns Official Journal of the International Society for Pediatric Neurosurgery. 2014; 30(3): 535-540.
47. Weiss J, Sauer A, Divac N, et al. Interaction of angiotensin receptor type 1 blockers with ATP-binding cassette transporters[J]. Biopharmaceutics & Drug Disposition. 2010; 31(2): 150-161.

Figures

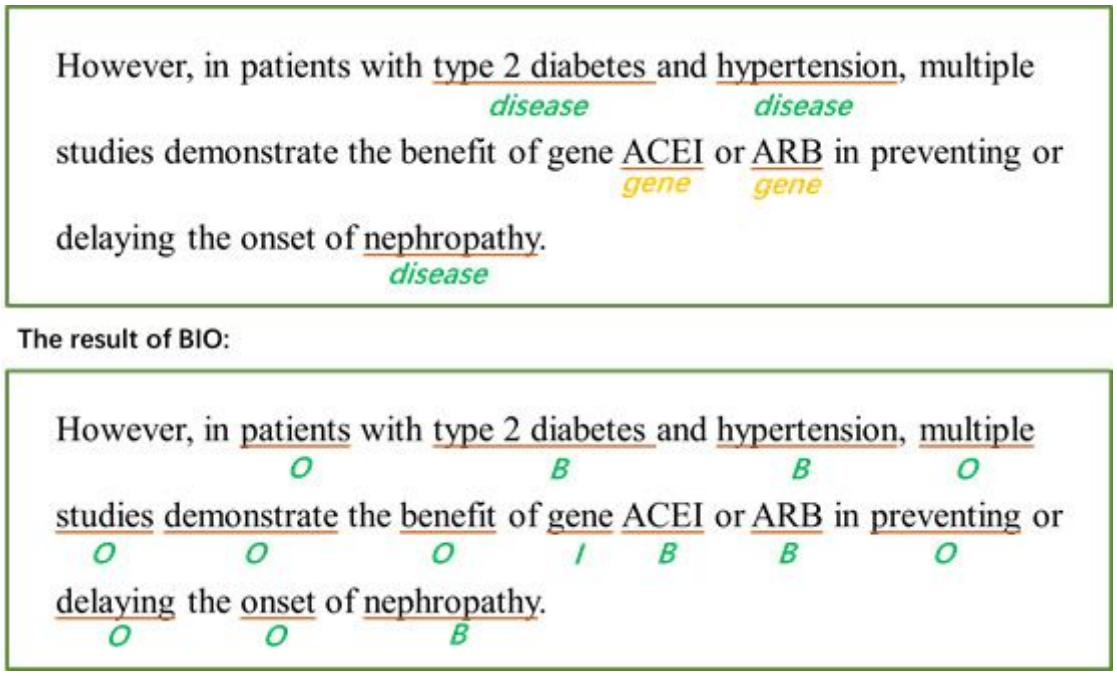

Figure 1

Example of boundary determination of bio-entity using BIO

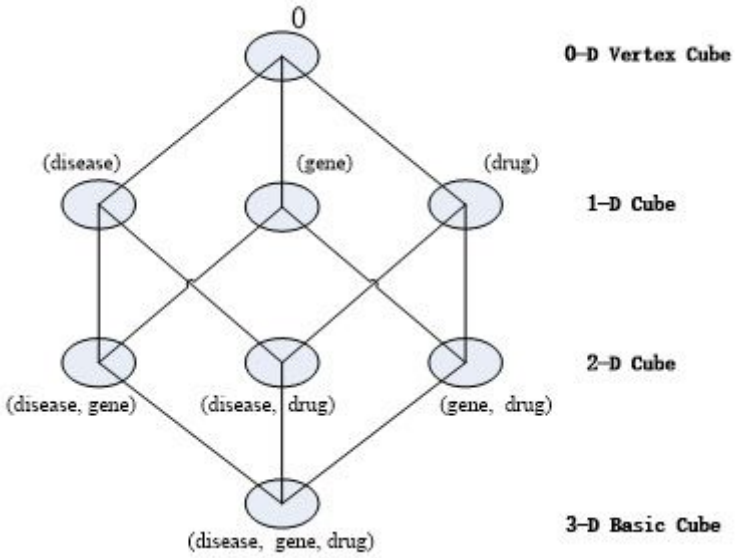

Figure 2

Sketch graph of the three-dimensional data cube. Each cube represents a different grouping. The basic cube contains three dimensions: disease, gene, and drug.

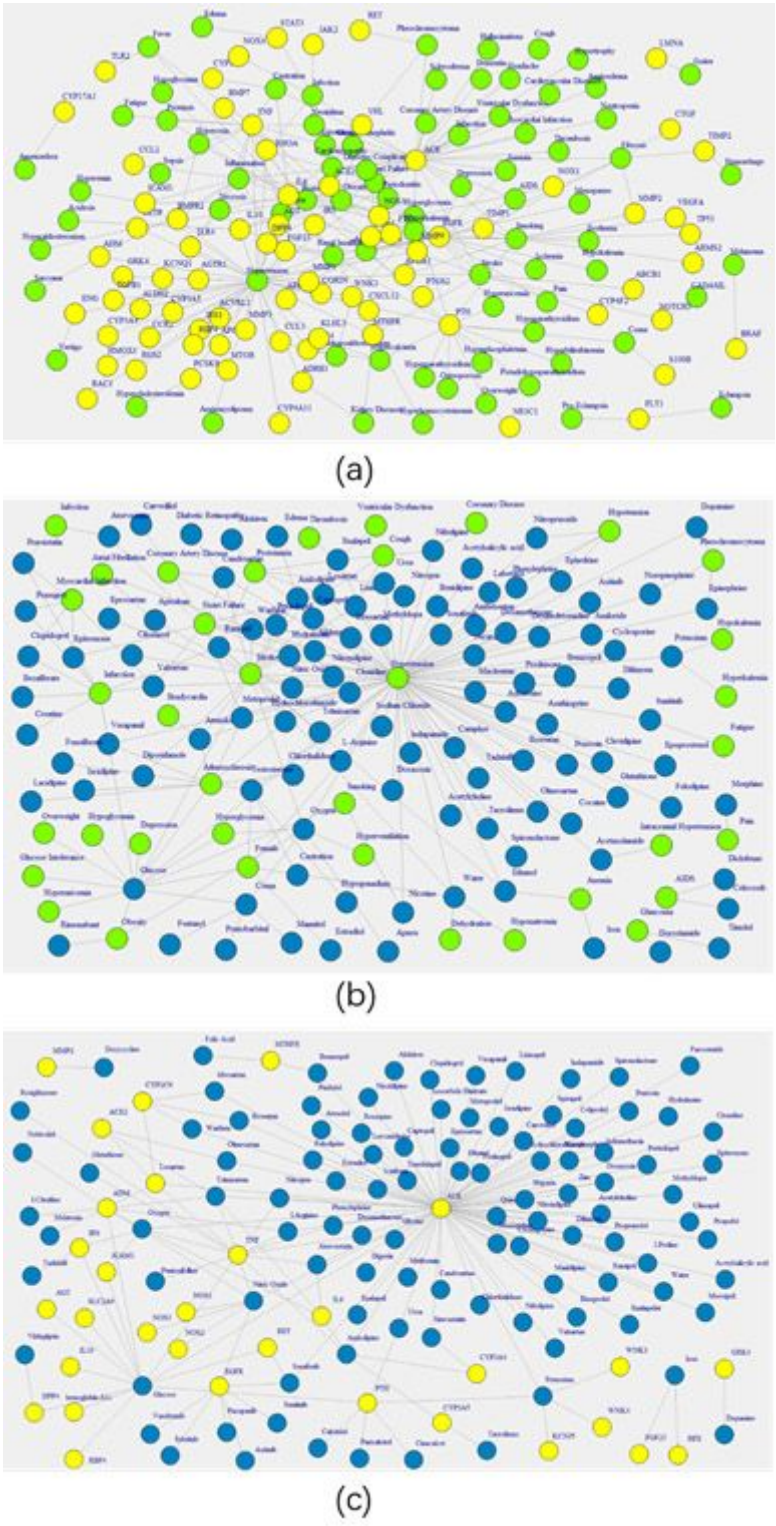

Figure 3

2-D disease-gene cube network of hypertension. (a) 2-D disease-gene cube; (b) 2-D disease-drug cube; (c) 2-D gene-drug cube. To each of the node there corresponds a bio-entity and to each cross-over there corresponds an association. The larger the lift value between nodes, the closer the distance. The green nodes represented diseases, the green nodes represented diseases, the blue nodes represented drugs, and the yellow nodes represented genes.

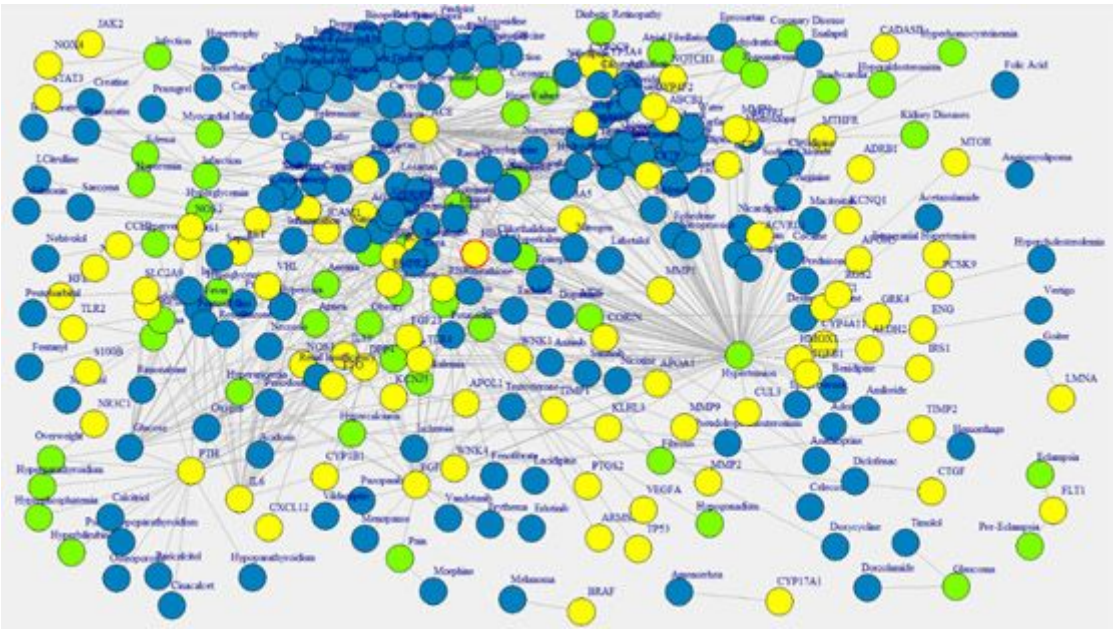

Figure 4

3-D Disease-gene-drug basic cube network.

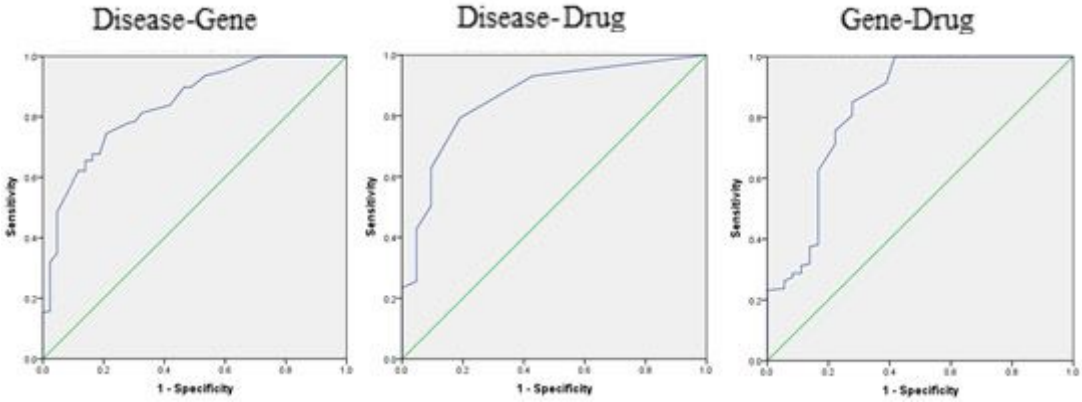

Figure 5

ROC curve evaluation. ROC curves are shown for disease-gene, disease-drug, and gene-drug predictive associations analyses for several intermediate inclusion criteria.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 1D.xlsx
- 2Dgenedrug.bmp
- 2Ddisedrug.bmp
- 3Ddisedruggene.bmp

- 2Ddisegene.bmp
- 3D.xlsx
- 2D.xlsx