Topology-Driven Trend Analysis for Drug Discovery

Yanhua Lv[1], Ying Ding[2,3], Min Song[4*], Zhiguang Duan[1,5]

[1] School of Management, Shanxi Medical University, 56 Xinjian South Road, Taiyuan, Shanxi, China.

[2] School of Informatics & Computing, Indiana University, 107 S. Indiana Ave, Bloomington, IN, USA.

[3] School of Information Management, Wuhan University, Wuhan, Hubei, China.

[4] Department of Library and Information Science, Yonsei University, Seoul, Korea. Email: min.song@yonsei.ac.kr. 82221232416

[5] Shanxi University of Chinese Medicine, 121 Daxue Street, Jinzhong, Shanxi, China.

* Corresponding author

Abstract

The primary goal of the present study is to discover new drug treatments by topology analysis of drug associations and their therapeutic group network. To this end, we collected 19,869 papers dated from 1946 to 2015 that are related to autism treatment from PubMed. We extracted 145 drugs based on MeSH terms and their synonyms (the total number is 6624) within the same ATC classification hierarchy and used them to find drug associations in the collected datasets. We introduced a new topology-driven method that incorporates various network analyses including co-word network, clique percolation, weak component, pathfinding-based analysis of therapeutic groups, and detection of important drug interaction within a clique. The present study showed that the in-depth analysis of the drug relationships extracted from the literature-based network sheds new light on drug discovery research. The results also suggested that certain drugs could be repurposed for autism treatment in the future. In particular, the results indicated that the discovered four drugs such as Tocilizumab, Tacrolimus, Prednisone, and Sulfisoxazole are worthy of further study in laboratory experiments with formal assessment of possible effects on symptoms, which may provide psychologists, physicians, and researchers with data-based scientific hypotheses in autism-drug discovery.

1. Introduction

Trajectory analysis of the research profile of successful therapeutic drugs is of paramount importance for the drug research community due to the following reasons. First, by analyzing research trends and thrusts of drugs, the evolution of the research profile of drugs can be understood, which in turn enables to scrutinize the impact of the initial use of the drugs in experimental studies on the subsequent studies of the drugs used for drug to humans. Second, inference of the new relationship between drugs through time-series analysis leads to a new drug discovery. However, despite the cruciality of such needs and potential synergetic contributions of informetics approaches, informetics has not been fully applied to solve the core problems of drug discovery. The majority of informetrics research in drug study dealt with measuring scholarly activities of drug discovery either by co-word or co-MeSH term analysis (Hong et al., 2016; Bordons et al., 2004; Leydesdorff et al., 2012). Otherwise, bibliometrics approaches to drug discovery were mainly focused on how particular public databases or resources have an impact on drug discovery (Cheng et al., 2014). To the best of our knowledge, there has been no previous study of applying the informetrics approach for solving new drug discovery. The proposed approach is the first attempt to identify new relationships among therapeutic groups of drugs and use them to infer new drug discovery for autism treatment by

mapping drug associations onto their therapeutic groups. Our informetrics-centric approach coupled with text mining makes it possible to make use of trajectory analysis of therapeutic group networks for new drug discovery.

As the subject for the present study, we chose autism. Despite the increasing diagnosis of autism over the last twenty years and extensive biomedical research on brain and nervous system disorders, and related pharmacological problems, there has been little progress in the development of pharmacological treatments for the social impairments that are at the core of this disorder (Modi and Young, 2012; Martin et al., 1999; FDA, 2009). Accordingly, Norén argued that data-driven discovery can be considered in drug repositioning because it can produce no pre-specified hypothesis (Norén, 2011). As asserted by Norén, the large amount of pharmacological and biological knowledge available in the literature makes it an increasingly feasible to find novel drug indications for existing drugs using an in silico approach. By combining network analysis with text mining, the goal of the present study is two-fold: 1) to identify research trends pertinent to autism treatments, by discovering novel interactions among different therapeutic groups in autism research; 2) to identify drug repurposing opportunities for autism and proposing possible new scientific hypotheses by novel interactions among drugs discovered from step 1.

A series of experiments showed that together with text mining techniques, our proposal of the topology-driven analysis was able to detect the new therapeutic groups of drugs and new, plausible drug discovery in autism treatment research. In particular, the present study revealed that 50% of research focused on drugs in the same therapeutic group in the early stage of autism research (prior to mid-2000s). However, this proportion decreased with time, and more than 70% of research focused on cross-therapeutic group drugs in the past 10 years. In addition, the results of the study identified that core therapeutic groups of autism have steadily changed over time. However, there is a stable subgroup existing in the drug network (i.e., that never changes with time in structure), which forms the foundation in autism treatment. In the drug network, we identified four drugs as worthy of further study in laboratory experiments, which may provide psychologists, physicians, and researchers with data-driven scientific hypotheses in autism-drug discovery.

2. Related Work

All drugs used in clinical medicine require large-scale trials before approval (Jeong et al., 2016), which provides a wealth of material in the published literature about the biological activities and safety of the drugs. However, the amount of such information is now too large for any one person to keep abreast of, even in a niche area of research. Scientific publications are growing at an exponential rate (Larsen and Von Ins, 2010), with over 50 million papers published so far (Jinha, 2010), and over a million additional articles published annually (Björk et al., 2008). That means on average a new article is published every 30 seconds. At the same time, digital publications are narrowing their science and scholarship focus, as well as the range of findings and ideas built upon them (Evans 2008). Increasing interest in improving treatments for autism also has led to a surge in publications in this field; far more than any individual scientist can keep up with. As a result, data mining and information technologies have proven necessary for good research-based decision-making (Chen et al., 2012; Bianchi et al., 2014). Articles have been essential to bibliometric studies for decades (Ying et al., 2013). Based on articles, the concept of Entitymetrics was proposed (health and news) to measure the impact of knowledge units at various levels. One kind of entitymetrics are micro-level knowledge entities, such as genes, drugs, and diseases etc., and they act as carriers of knowledge units in scientific articles to identify the importance of entities embedded in scientific literature for

further knowledge discovery. In the present research, literature-based knowledge discovery aims to connect the potential relationships of scientific entities to generate new knowledge from the perspective of drugs.

Secondly, as medicine advances, all approved drugs must be novel and important to be reasonably characterized as cures in their own domains. However, drug research and development requires huge investments, which means that an unclear target might result in additional costs or inefficiency (Williams et al., 2015). From this perspective, a scientific hypothesis can facilitate medical trials, as well as providing substantial data for the advance of science and improvement of medicine. Although experiment-based knowledge discovery is based on stringently validated data from experiments or clinical trials, the benefits of literature-based discovery can be enormous in helping domain experts to form scientific hypotheses. The connections between concepts in scientific literature can be established if two concepts co-occur in a predefined context (e.g., title, abstract, one sentence, or one paragraph), and researchers had verified its valuation. Stegmann and Grohmann also argued that co-occurrence network analysis is a powerful method for literature-based hypothesis generation and knowledge discovery by finding characteristic values in the co-keyword analysis which allow a rapid identification of possible cluster based on centrality-density ratio (Stegmann and Grohmann, 2003). After that, a study shows that Metformin changes the peroxisome proliferator-activated receptor in the uterine tissue of mice (Blumberg et al., 2013), and then researchers identified the interaction between Metformin and the peroxisome proliferator-activated receptor through a bio-entity citation network. In 2013, a research identified of an interaction between Metformin and Resistin, which supported (Newschaffer et al., 2007) the hypothesis that Metformin treatment had a positive impact on up-regulating Resistin.

Given the rapid growth of scientific literature, literature-based approaches to generating hypotheses automatically have gained increasing attention in recent years. Swanson discovered that certain unintended logical connections across scientific domains and potentially revealing of new knowledge, were enabled by reference citations or other bibliographic clues (Swanson, 1987). By weaving the related but disjoint literatures together, implicit, unnoticed hypotheses can be generated (Swanson, 1987). Swanson pioneered literature-based discovery to mine potential and valuable relationships among biological concepts from public knowledge (Swanson, 1986; Swanson, 1988; Swanson and Smallheiser, 1999). Swanson pointed out that if the connections between literature A and B as well as literature B and C are known while the interaction between literature A and C is unknown, there is a potential, novel connection between literature A and C for new knowledge discovery. Employing this model, many disease-related hypotheses were proposed, including the relationship between fish oil and Raynaud's syndrome (Swanson, 1986), magnesium and migraine (Swanson, 1988), indomethacin and Alzheimer's disease (Smalheiser and Swanson, 1996), and metabolites and biological processes (Baek et al., 2017). By applying literature-based discovery to trend analysis of genes, the combination of bibliometrics and data-mining techniques based on MeSH was proposed to detect "hot" trends of emerging molecular mechanisms for obesity, which leads to discover emerging areas of scientific research within obesity (Rajpal et al., 2011). A literature-based trend analysis was proposed to explore potential biological connections between gene and disease, which can be utilized for drug discovery (Rajpal et al., 2014). By considering multiple ontologies and biological databases, a literature-based analysis was proposed to find detailed relationships between drugs, proteins and diseases for drug-repositioning discovery (Wei et al., 2014). In addition, researchers refined and extended ABC model to discover hidden connections (Gordon and Lindsay, 1996; Lindsay and Gordon, 1999). Chen and his colleagues proposed an explanatory and computational theory of scientific discovery by extending the concept of structural holes from

social networks to co-citation and collaboration networks (Chen et al., 2009). There was a recent attempt to discover potential relationships among biomedical concepts in biomedical literatures by a storytelling-based semantic path analysis (Song et al., 2015). Furthermore, some researchers have developed prediction approaches based on existing links in social networking or datasets, in which it is assumed that two nodes are similar if they share similar topology (e.g., a certain number of neighbors, and similar shortest paths) (Chen et al., 2012; FDA, 2012; He, 1999), and proposed a pathfinding and statistical model (Chen et al., 2012). Kostoff adopted a citation-based linkage through bibliographic coupling for literature-related discovery (Kostoff, 2014). Small and his colleagues identified biomedical discoveries using citation contexts (Small et al., 2017). In the context of drug discovery, drug-target interactions can be predicted based on observed topological features of a semantic network across the chemical and biological space. A new framework was proposed to predict drug-target interactions using a semantic network that integrates chemical, pharmacological, genomic, biological, functional, and biomedical information into a unified framework (Nooy et al., 2005). It offers the flexibility to enrich the feature space by using different normalization processes on topological features, and it can perform model construction and feature selection at the same time.

3. Methods

3.1. Data collection

All the autism-related publications were retrieved from PubMed. We collected 19,869 papers dated from 1946 to 2015, including research articles, letters, reviews, research support, and others, using the following 9 Medical Subject Headings (MeSH) terms relevant to autism: Autistic disorder; Autism spectrum disorder; Rett syndrome; Akinetic mutism; Macrocephaly autism syndrome [supplementary concept]; AUTS2 protein, human [supplementary concept]; Auts2 protein, mouse [supplementary concept]; Adenylosuccinate lyase deficiency [Supplementary Concept]; GoPro49 protein, human [Supplementary Concept]. These nine MeSH terms were obtained by querying PubMed with the term "autism" for the MeSH field search. A total of 251 drugs were extracted from the Food and Drug Administration (FDA) drug directory using MeSH (version: 5/24/2016) (FDA 2012). However, since several drug names are ambiguous, we decided to exclude those drugs from analysis and three physicians were involved in filtering the list which resulted in 145 drugs, including behavior (9 terms), disease names (13 terms), symptom (24 terms), and technical terms (58 terms). Those excluded terms are listed in Table 1. These were classified using the Anatomical Therapeutic Chemical (ATC) classification system.

Table 1 Drug list screened and kept for drug association analysis

Abandoned terms (106)	*Remaining terms (145)*
Guilt; Anger; Digestion; Drainage; Metabolism; Solutions; Therapeutic; Weight Loss; Speech Therapy; Alcoholism; Arthritis; Cerebral Palsy; Colitis; Psoriasis; Sinusitis; Tonsillitis; Whooping Cough; Herpes Simplex; Hypertension; Hypoglycemia; Asthma; Burns; Headache; Infection; Inflammation; Injuries; Anxiety;	Acetaminophen; Acetic Acid; Acetylcysteine; Adenosine; Allopurinol; Alprostadil; Amoxapine; Amphotericin B; Aripiprazole; Ascorbic Acid; Aspirin; Norepinephrine; Oxytocin; Paclitaxel; Paroxetine; Pentoxifylline; Phenytoin; Pimozide; Potassium Chloride; Prednisolone; Prednisone; Probenecid; Estrogens; Etomidate; Famotidine; Fentanyl; Flumazenil; Fluorouracil; Fluoxetine; Folic Acid; Galantamine; Glucagon; Glutathione; Dexmedetomidine; Dextroamphetamine; Diazepam; Digoxin; Dinoprostone;

Constipation; Cough; Depression; Diarrhea; Dyspepsia; Eczema; Edema; Enuresis; Fatigue; Fibromyalgia; Flatulence; Nausea; Neuralgia; Pain; Tinnitus; Wound Healing; Grief; Hypersensitivity; Macular Degeneration; Foot; Gold; Hand; Liver; Food Additives; Formaldehyde; Health; Hearing; Bacteria; Calcium; Calcium Carbonate; Candida; Candida albicans; Carbon Dioxide; Cholesterol; Collagen; Corpus Callosum; DNA; Ear; Female; Menopause; Metals; Nitrous Oxide; Nose; Oxygen; Phosphorus; Radiation; Selenium; Skin Care; Sleep; Spleen; Stomach; Sucrose; Sulfur; Urea; Vaccines; Water; Xenon; Zinc; Graphite; Hypothalamus; Iris; Libido; Manganese; Ammonia; Aspergillus niger; Caffeine; Chromium; Copper; Phenylalanine; Borrelia burgdorferi; Tryptophan; Tyramine; Ginkgo biloba; Glycine; Growth Hormone; Histamine; Hypericum; Lactic Acid; Turkey.	Diphenhydramine; Dopamine; Droperidol; Ephedrine; Epinephrine; Estradiol; Topotecan; Trazodone; Tretinoin; Trifluoperazine; Valproic Acid; Vancomycin; Venlafaxine Hydrochloride; Verapamil; Vitamin B Complex; Vitamin D; Zinc Sulfate; Atomoxetine Hydrochloride; Azacitidine; Azathioprine; Baclofen; BCG Vaccine; Betamethasone; Bleomycin; Bumetanide; Buspirone; Calcitriol; Carbamazepine; Prochlorperazine; Progesterone; Promethazine; Propofol; Propranolol; Quetiapine Fumarate; Reserpine; Riluzole; Risperidone; Rivastigmine; Saccharomyces; Guanfacine; Haloperidol; Hydrocortisone; Hydrogen Peroxide; Hydroxyzine ;Ibuprofen; Iodine; Lactulose; Leucovorin; Lidocaine; Lithium; Carbidopa; Celecoxib; Cetirizine; Chlordiazepoxide; Chlorogenic Acid; Chlorpromazine; Citalopram; Clindamycin; Clonazepam; Clonidine; Clozapine; Saccharomyces cerevisiae; Salicylic Acid; Scopolamine Hydrobromide; Serotonin; Sertraline; Simvastatin; Sincalide; Sirolimus; Sodium Bicarbonate; Sodium Chloride; Spironolactone; Lithium Carbonate; Lorazepam; Lovastatin; Loxapine; Magnesium Oxide; Magnesium Sulfate; Mannitol; Melatonin; Memantine; Methotrexate; Methylphenidate; Colchicine; Cromolyn Sodium; Cyclophosphamide; Cycloserine; Cyclosporine; Cyproheptadine; Dacarbazine; Dexamethasone; Sulindac; Sumatriptan; Tacrolimus; Taurine; Testosterone; Theophylline; Thiamine; Thiothixene; Metoprolol; Midazolam; Minocycline; Misoprostol; Naproxen; Niacin; Nicotine; Nifedipine.

3.2. Proposed Approach

Fig 1 illustrates how the proposed work was conducted. The overall process of the proposed approach consists of the following major steps.

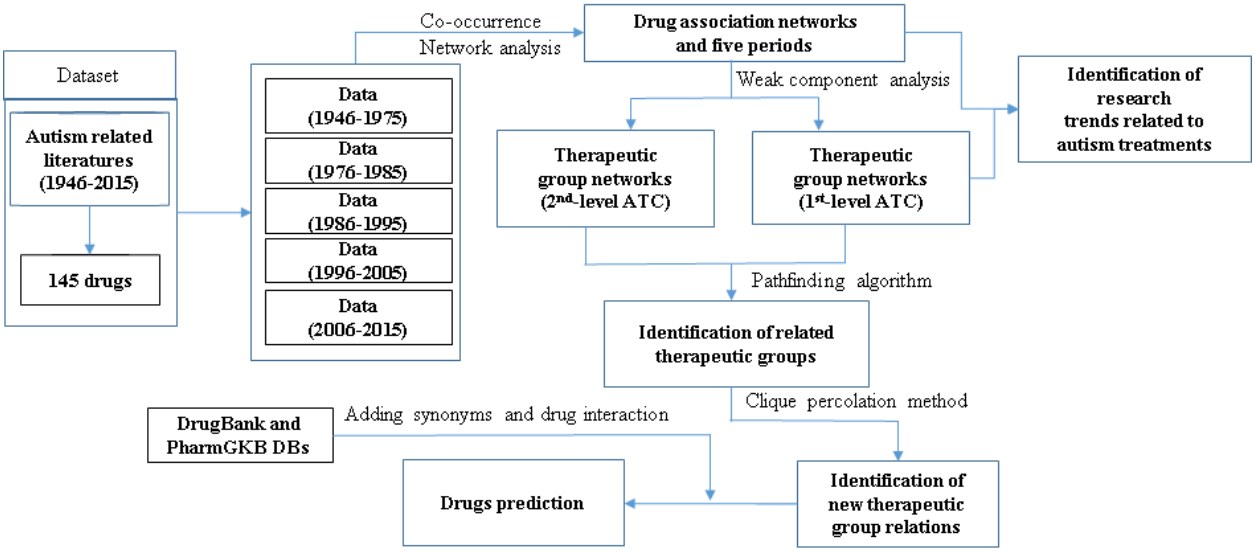

Fig 1 Overall Architecture of the Proposed Approach

Step 1: Using the dataset described above, with either a 1st-level or 2nd-level ATC classification code, we built the six co-occurrence drug networks including one overall and five networks for five periods.

Step 2: We identified relationships and communities among the drugs from the networks built per period to examine how these change over time. To this end, we employed three different techniques for further analysis. First, we adopted weak component analysis to identify similar drug associations in the network.

Step 3: Second, we applied the pathfinding algorithm to find potential relationships among the therapeutic groups resulted from Step 2. The analysis was based on the previous four periods, and the potential relations were detected based on the 5th period.

Step 4: Third, by applying CPM to extract Maximal-cliques from the network by 2nd –level ATC classification code, we proposed drugs that might be considered candidates for repurposing for Autism. These candidates, along with the literature analysis, can in turn be used to investigate new scientific hypotheses.

Co-occurrence network analysis: This technique was developed during the 1980s and has been employed to map the dynamic development of several research fields (He, 1999; Bauin et al., 1991). Based on the co-occurring frequency of every two words in a PubMed record, the major themes of a given therapeutic group can be identified. The usefulness of co-word analysis techniques has been proven in determining the extent to which these strategies contribute to the definition of the thematic structure of a research network at any given moment in time (Bauin et al., 1991). In this paper, it was established to identify the importance of the drugs in the network by their centralities. We built several different networks represented by different entities such as drug, drug interaction, and hierarchical ATC classification codes at a different level. We then employed weak component analysis to find drug associations in the network. A weak component is a maximal weakly connected subnetwork in which all nodes are connected by

a semi-path (Nooy et al., 2005). The word "maximal" means that no other node can be added to the subnetwork without destroying its defining connectedness.

Pathfinding-based detection of related therapeutic groups: We adopted a pathfinding-based detection technique to identify the potential relationships between therapeutic groups on the built network. The theoretical basis is that if there is a relation between node A and B and between B and C, then the relation between A and C is regarded as a potential relation. A pair of nodes with one or more common neighbors becomes a path candidate. Here, we used a heap-based Dijkstra algorithm to quickly find paths between two nodes (Wang et al., 2011). The Dijkstra algorithm for pathfinding achieves a complexity of O ($n\log n$), which is relatively faster than other comparative algorithms (Chen et al., 2012). We selected only paths whose length of a path is greater than one and pairs of nodes that connect one or more common neighbors with each other. If the identified node pair (the drug pair) is found in the next period network in networks partitioned by certain periods, it is likely to be regarded as potentially valuable autism treatment, and only those drugs are fed into the pathfinding method for therapeutic group analysis.

The pathfinding is described below:

Let there be a graph G (V,E), with $P_l(s \to t)$ as the lth-shortest path from node s to node t; $e_{i\to j}$ as the edge from node i to node j; and R_{ij} as the link type of e_{ij}. Then, the probability of traversing from s to t via a path is (Chen et al., 2012):

$$p\big(P_l(s \to t)\big)=\prod_{i=1}^{m-1} p(e_{i\to i+1}) \quad (1)$$

An undirected graph consists of two parts of a path: from s to t, and from t to s. Thus, in an undirected graph, the algorithm can be described as (Chen et al., 2012):

$$\log(\, p\big(P_l(s,t)\big))=\frac{(\log(\, p\big(P_l(s\to t)\big))+\log(\, p\big(P_l(t\to s)\big)))}{2}. \quad (2)$$

Clique percolation method (CPM): We adopted CPM, a community detection technique, to analyze maximal-cliques to detect associations between therapeutic groups mapped from drugs. CPM allows overlaps between communities in a natural way. If a certain sub-graph fulfills the criteria to be considered a community, then it remains in a community, independent of what happens to any other part of the network. For example, the CPM has been used to detect communities in the studies of cancer metastasis through various social networks (Jonsson and Bates, 2011; Jonsson et al., 2006; Palla et al., 2007; Toivonen et al., 2006).

Clique-based graph clustering and finding important drugs: Since there is no common drug found in a clique, we expanded the drug list by adding synonyms of a drug to the clique to which it belongs. For synonym expansion, we used DrugBank and PharmGKB. The total number of synonyms is 6,624. The number of synonyms per ATC is listed in Appendix 1. Subsequently, for 34 cliques in Table 5, we paired up drugs listed in each clique. This means we built an adjacency matrix per clique. We then computed a PageRank score of each drug and show top-10-ranked drugs, which can be treated as important drugs for the clique. Another approach to detecting important drugs is by interactions between drugs within a clique. To this end, we used the drug interaction list for a drug available in DrugBank. For a drug in a clique, we only considered drugs interacting with the drug if an interacting drug is found in the same clique. Since there are multiple drugs interacting with a drug, we paired up interacting drugs to create an adjacency matrix. Upon the built matrix, we computed the PageRank score of a drug to

identify important drugs interacting with the target drug within a clique. We displayed top-10-ranked drugs by PageRank.

4. Results and Discussion

All of the 145 drugs and their synonymous drug names (6,624 synonyms) are classified according to the ATC classification system, in which active substances were divided into different groups according to organ or system on which they act, and their therapeutic, pharmacological, and chemical properties. According to this classification, drugs are grouped into fourteen basic groups (1st level) according to the organic system of the organism where they work. These fourteen anatomical groups represent the first anatomical level and are labeled with one capital letter. Drugs are further divided into therapeutic groups (2nd level), which are often used to identify pharmacological subgroups when that is considered more appropriate than therapeutic or chemical subgroups. 1st and 2nd level classification codes are used in this research (1L and 2L, respectively, for short).

4.1. Analysis of Drug Associations

Based on co-occurrence of drugs, we built the drug association networks where each node denotes a drug and colored each node with its therapeutic group according to the 1st-level ATC classification code of drugs (Fig 2). In other words, drugs assigned the same 1st-level ATC code are painted the same color. Out of all 1st-level ATC classification codes, there are 13 therapeutic groups involved in autism research, and these therapeutic groups and their corresponding alphabetic codes are listed in Table 3.

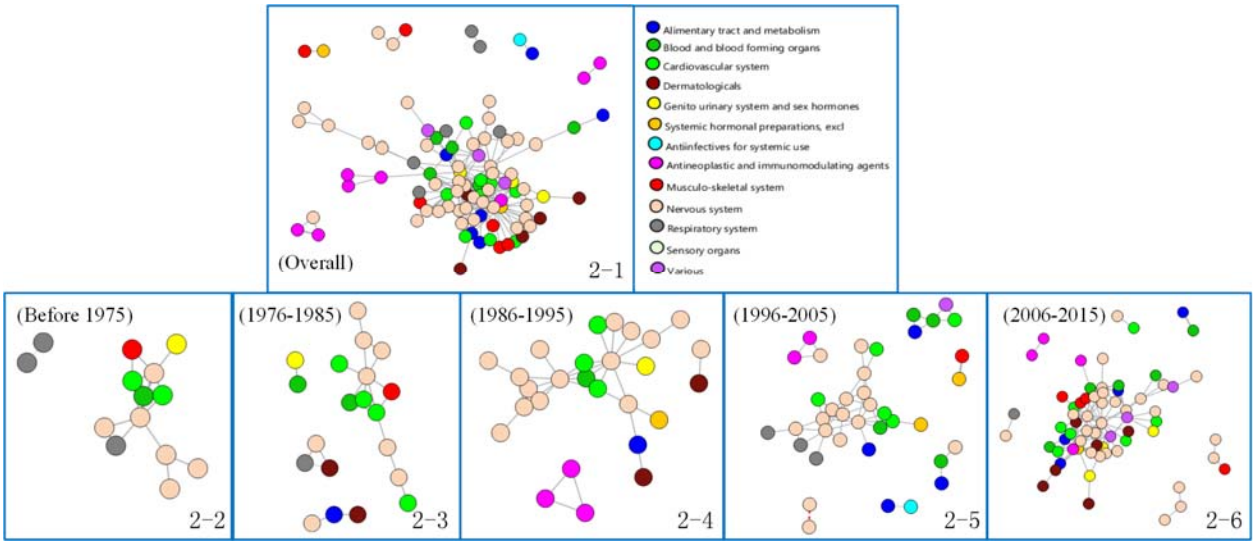

Fig 2 Drug association analysis of overall network and five periods' networks

To understand the change of drug research over time, we divided the dataset into five periods: before 1975, 1976–1985, 1986–1995, 1996–2005, and 2006–2015, labelled with 2-2, 2-3, 2-4, 2-5, and 2-6, respectively. Fig 2-1 illustrated the overall drug association network. These six networks were built and analyzed in the following manner: First, drug association was defined to show groups formed according to their relationships in the drug network. We detected drug associations on the basis of drug pairs from the drug network. Second, drug associations were examined by weak component analysis. Drugs in the same component were represented by ATC classification

codes where nodes in the same therapeutic group are painted the same color. In Fig 2-1, seven groups were identified, in which only two groups have the same color, and the others have more than one color, which is also shown in Table 2. This indicates that drugs coming from different therapeutic groups were relevant to autism research.

Overall, Fig 2-1 to 2-6, also reflected in Table 2, show that in the early stage, about 50% of research was focused on drugs in the same therapeutic group in treating autism, and the proportion decreased with time, which indicates that the cross-therapeutic group research was more widely pervasive recently than ever before. During the latest stage, more than 70% of drug research was focused on cross-therapeutic group drugs, which was computed based on the second column and fourth column of Table 2 for the period of 2006-2015 (5/7=71.4%). From the perspective of drug research scale, the maximum number of early-stage drugs was only 12 in six therapeutic groups. However, as the evolution of autism treatments, 42 drugs appear in one association in 10 therapeutic groups. This in turn shows the more active interaction between different therapeutic groups including Alimentary tract and metabolism; Blood and blood-forming organs; Cardiovascular system; Dermatological; Genito-urinary system and sex hormones; Antineoplastic and immunomodulating agents; Musculo-skeletal system; Nervous system; Respiratory system; and Various (Table 3).

Table 2 Statistics of drug associations in all periods.

Periods	*Number of groups*	*Number of groups with the same color*	*Number of groups with different colors*	*Number of nodes in the biggest group*	*Number of colors in the biggest group*
Before1975	2	1	1	12	6
1976–1985	4	0	4	12	4
1986–1995	3	1	2	22	7
1996–2005	7	1	6	22	6
2006–2015	7	2	5	42	10
Overall network	7	2	5	82	11

Based on the six drug networks in Fig 2, we summarized the statistics of the autism-related drug associations in each therapeutic group and their proportion against the total drug number in each therapeutic group in Table 3. By the numbers and proportions of drugs, the therapeutic group, nervous system, was ranked no. 1 in all periods, followed by the therapeutic groups, cardiovascular system and blood and blood-forming organs, followed while their proportions tended to increase steadily. Furthermore, it is worth noting that the number of drugs from the dermatological, the genito-urinary system, and sex hormones therapeutic groups also increased in autism research. On the other hand, drugs for therapeutic group, the respiratory system, decreased with time.

Table 3 Statistics of drug associations in all therapeutic groups

Code	*Therapeutic group name*	*Fig3-1*		*Fig3-2*		*Fig3-3*		*Fig3-4*		*Fig3-5*		*Fig3-6*	
		N_n	P_n	N_n	P_n	N_n	P_n	N_n	P_n	N_n	P_n	N_n	P_n
A	Alimentary tract and metabolism	6	6.25%	0	0	1	5.00%	1	3.70%	4	10.26%	4	6.06%

B	Blood and blood forming organs	5	5.21%	1	7.14%	2	10.00%	1	3.70%	4	10.26%	5	7.58%
C	Cardiovascular system	9	*9.38%*	2	14.29%	4	20.00%	3	11.11%	5	*12.82%*	7	*10.61%*
D	Dermatologicals	5	5.21%	0	0	1	5.00%	1	3.70%	0	0	5	7.58%
G	Genito-urinary system and sex hormones	3	3.13%	1	7.14%	1	5.00%	1	3.70%	0	0	3	4.55%
H	Systemic hormonal preparations, excl	2	2.08%	0	0	0	0	1	3.70%	2	5.13%	1	1.52%
J	Antiinfectives for systemic use	1	1.04%	0	0	0	0	0	0	1	2.56%	0	0
L	Antineoplastic and immunomodulating agents	8	*8.33%*	0	0	0	0	3	*11.11%*	2	5.13%	4	6.06%
M	Musculo-skeletal system	6	6.25%	1	7.14%	1	5.00%	0	*0*	1	2.56%	4	6.06%
N	*Nervous system*	*42*	*43.75%*	*6*	*42.86%*	*9*	*45.00%*	*16*	*59.26%*	*16*	*41.03%*	*29*	*43.94%*
R	Respiratory system	6	6.25%	3	*21.43%*	1	5.00%	0	0	3	7.69%	1	1.52%
S	Sensory organs	0	0	0	0	0	0	0	0	0	0	0	0
V	Various	3	3.13%	0	0	0	0	0	0	1	2.56%	3	4.55%

Notes. N_n represents the number in a certain therapeutic group; P_n represents the proportion in the total number of a certain therapeutic group.

Examining the drug research of the last stage shown in the last column of Table3, we see that drugs for the therapeutic groups like nervous system, cardiovascular system, and dermatology, blood and blood forming organs, occupied the top four positions, which indicates that drugs from these therapeutic groups were popular during that period. On the other hand, drugs for the therapeutic group, respiratory system and systemic hormonal preparations, decreased in proportion over the same period.

4.2. Therapeutic Group Analysis in Autism Research

In this sub-section, we aim to grasp the core structure of autism-related therapeutic groups by mapping drugs to their corresponding 1st-level ATC classification codes and making therapeutic group associations. We called the built network the “therapeutic group network” because ATC classification divided drugs into different groups according their therapeutic and chemical characteristics such as Cardiovascular system and Dermatologicals, which is similar with MeSH. Each node in each network is a therapeutic group in the ATC classification, which represents a therapeutic group related to autism. There are 13 therapeutic groups involved in autism research.

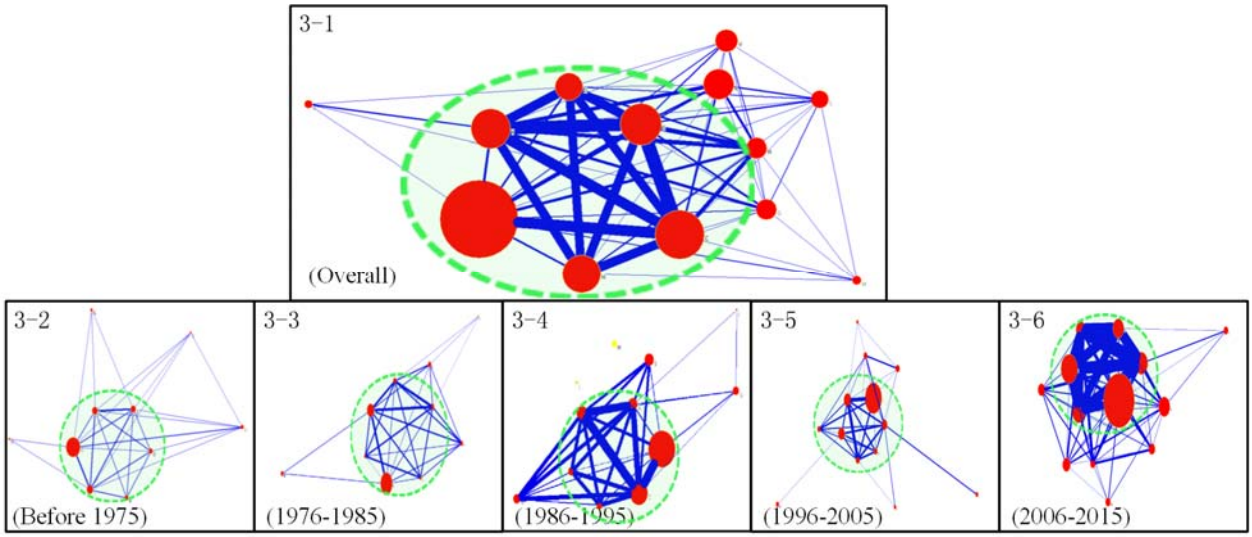

Fig 3 Therapeutic group relationships over time and in five periods

The six boxes of Fig 4 represent six therapeutic group networks. Each node in the network is a therapeutic group represented by the ATC classification (therapeutic group) in autism. The upper box of Fig 3 (Fig 3-1) shows the relationships among the 13 therapeutic groups by weak component analysis whereas the five lower boxes represent the five stages of therapeutic group relationships.

In Fig 3-1, the biggest node is the therapeutic group, nervous system, followed by cardiovascular system. A subgroup within the green colored dot line consists of six nodes with strong edges among them. In this group, therapeutic groups such as "N (Nervous system)," "C (Cardiovascular system)," "H (Systemic hormonal preparations, excl)," "S (Sensory organs)," "D (Dermatologicals)," and "A (Alimentary tract and metabolism)" are closely related in the autism treatments domain ("6CN" for short). In terms of topology analysis, this is a complete subnetwork (the strictest structural form of a cohesive subgroup, which is called a clique: a set of vertices in which each vertex is directly connected to all other vertices). From the perspective of medication treatment, they represent components of the core set-up usually combined to treat the basic symptoms of autism.

The five stages shown in Fig 3-2 to 3-6 enable to provide the trajectory view of the medication treatment of autism from 1957 to 2015. There are 10 nodes in Fig 3-3, which is the largest network, and the subgroup 6CN mentioned above is shown clearly. This indicates that the relationships among the six nodes of 6CN were intensively studied during that period. In addition, a medication treatment structure was created since that time, and is the primary medication approach to treating autism. In Fig 3-3, the therapeutic group, N (Nervous system), is connected with other nodes weakly, but forms a complete network. This implies that nervous system symptoms are generally considered to be very important in treating autism. The therapeutic group, R (Respiratory system), is closely connected with 6CN, and a new complete subgroup (including 7 therapeutic groups) has emerged, indicating that treating respiratory-system symptoms was also considered important in treating autism over the indicated period, and was often combined with the core set-up.

Unlike in Fig 3-3, therapeutic groups, J (Antiinfectives for systemic use) and L (Antineoplastic and immunomodulating agents), appear in Fig 3-4. This implies the beginning of treating autism from the perspective of anti-infectives and antineoplastic and immunomodulation medications, and it broadens the approach to autism

treatment. However, antineoplastic and immunomodulation medications were used individually between 1986 and 1995. During 1996 to 2005, 13 therapeutic groups evolved into one network.

Fig 3-6 illustrates the interaction research among the 6CN therapeutic groups, which can be seen more clearly by thick lines than before, although there are numbers of therapeutic groups involved in its parent groups. As it is always sharp-edged in the six 6CN networks and has the strong connection in all the periods, we see their central position in autism research and treatment in all cases and periods. At the same time, in the last period, therapeutic groups, R (Respiratory system), G (Genito-urinary system and sex hormones), and B (Blood and blood forming organs), were connected to 6CN completely, which shows the broadened scope of the research by exploring symptoms from the various therapeutic groups including Respiration, Genito-urinary and Sex hormones, and Blood systems in autism.

4.3. Discovering New Therapeutic Group Relations in Autism Treatments

Based on the six therapeutic group networks described earlier, we identified all of the common neighbors and their characteristics in the four networks (before 1975, 1976–1985, 1986–1995, and 1996–2005) by the pathfinding algorithm and CPM. We first applied the pathfinding algorithm to extract the related therapeutic groups that were later fed into CPM as input to detect cliques. Specifically, with a set of therapeutic groups represented by the 1st-level ATC classification code, CPM predicts that five sets of node pairs with common neighbors (>=1) representing five sets of therapeutic group relationships are likely to appear in the next period. Subsequently, we verified every set of node pairs identified above in the later networks (1976–1985, 1986–1995, 1996–2005, and 2006–2015). By excluding the existing relations in the previous periods, we were able to find where the new relationship emerges. The results are shown in Table 4.

Table 4 Verifying result of therapeutic group relationships at the 1st level

Prediction based on the network	*Number of predicting node pairs*	*Number of node pairs with common neighbors (>=1)*	*Number of common neighbors (Total/Min:Max)*	Verifying in the network				Evaluation of predicting ($N_{\text{node pairs found}}/N_{\text{all node pairs}}$)
				D(1976-1985)	*D(1986-1995)*	*D(1996-2005)*	*D(2006-2015)*	
P(before1975)	0	0	/	/	/	/	/	0
P(1976-1985)	13	9	24(5:2)	/	C/G	H/M	B/D; B/H; C/G; D/G; G/H; G/R; M/R	69%
P(1986-1995)	13	7	25(5:1)	/	/	/	B/D; B/H; D/G; D/J; G/H; G/R; J/N	54%
P(1996-2005)	33	16	40(6:1)	/	/	/	A/V; B/D; B/H; D/G; D/J; D/V; G/H; G/L; G/R;	48%

							H/L; J/N; L/S; M/R; N/V; S/V; R/V	

There was no node pair (common neighbors >=1) of P (before 1975) found in the later periods. Ten pairs of P (1976–1985) were found in the later periods, one in the period of 1986–1995, one in 1996–2005, and eight in the last ten years. That means that in predicting 13 therapeutic group pairs, the previously nonexistent relationships between C (Cardiovascular system) and G (Genito-urinary system and sex hormones) appeared in the following period; the relationship between H and M appeared in 1996–2005; and the remaining eight pairs appeared in the last ten years. As 13 node pairs were predicted in this stage, and only ten pairs appeared, prediction accuracy for this period 69%.

In the same way, the evaluation in P (1986–1995) is 54%, and that of P (1996–2005) is 48%. Moreover, most of the predicted therapeutic group pairs appeared in the last ten years, demonstrating the booming development of autism research after 2006, and the explosive growth in multi-disciplinary studies among all these therapeutic groups in autism research. In general, the average prediction accuracy is 42.75%. As to the value of common neighbor numbers, there is no relationship between this value and evaluation, which means that the value does not affect prediction.

4.4. Discovering New Drug for Autism Treatments

The results based on the 1st-level classification only suggested a broad drug predisposition across different therapeutic groups. However, the issue of how to articulate a specific drug hypothesis with regard to autism research was yet to be resolved. Thus, we attempted to discover new, plausible drugs for autism treatment by mapping drugs to the 2nd–level ATC classification code and using the pathfinding algorithm and CPM used for finding new therapeutic group relations in autism treatments. The reason for using 2nd-level ATC classification code instead of 1st-level one is because the 1st -level ATC code is too broad to discover specific drugs commonly targeted by therapeutic groups. Fig 4-1 to 4-5 show the resulting networks of the five (1-5) cliques drawn from 1st-level ATC codes (1L1–5), respectively.

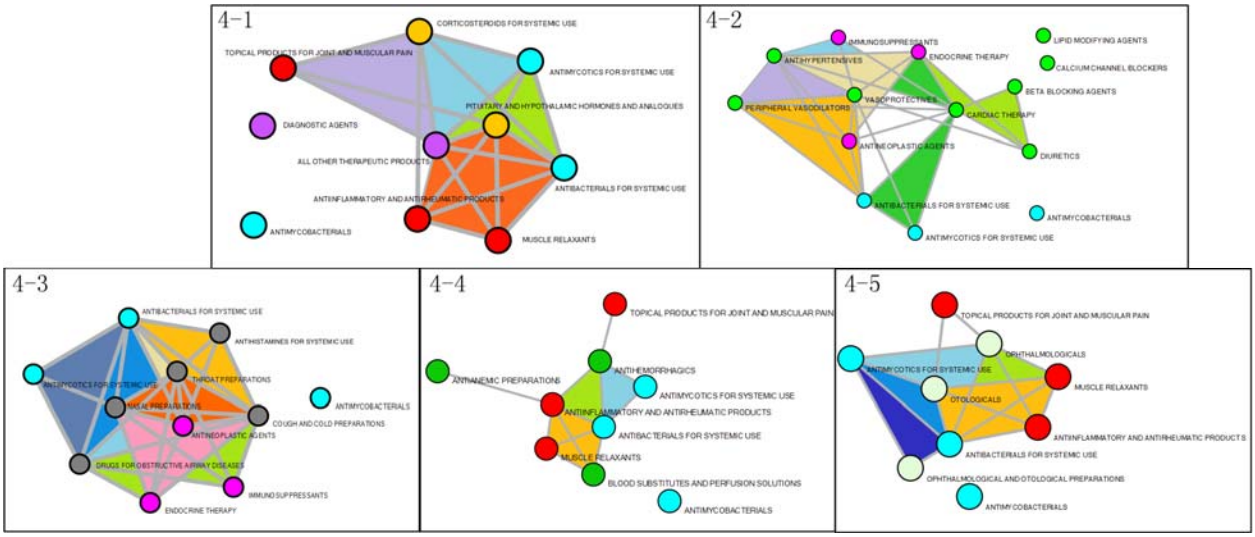

Fig 4 Cliques extracted from 1L1–1L5

Fig 4-1 to 4-5 show that 34 cliques were identified (in the five networks) and labeled in different colors. The areas covered by a clique are shared drugs, and the more nodes a clique includes, the more concrete the group is. The cliques identified at this stage are all 3 or more cliques. Those cliques are all very concrete drug groups that can be regarded as candidates of new drug discovery for autism research and treatment.

Since there is no common drug found in a clique, we decide to expand the drug list by adding synonyms of a drug to the clique to which it belongs. For synonym expansion, we use DrugBank and PharmGKB. Subsequently, for the 34 cliques mentioned above, we pair up drugs listed in each clique by building an adjacency matrix per clique. We compute a PageRank score of each drug and extract the top-10-ranked drugs, which can be treated as important drugs for the clique. Table 5 shows the common important drugs identified by expanded drug lists for a clique.

The important drugs for autism treatment were detected by drug interaction within a clique, which means they are the most relevant drugs for autism from the perspective of drug interaction. We used the drug interaction list for a drug available in DrugBank. For a drug in a clique, we only consider drugs interacting with the drug if an interacting drug is found in the same clique. Since there are multiple drugs interacting with a drug, we pair up interacting drugs to create an adjacency matrix. Upon the built matrix, we compute the PageRank score of a drug to identify important drugs interacting with the target drug within a clique. We extract the top 10 ranked drugs by PageRank, which results in 340 drugs (Table 5 and Appendix 2). Out of these 340 drugs, there are four drugs (Prednisone, Tocilizumab, Tacrolimus, and Sulfisoxazole) that are also found in the drug interaction list. We treat those four drugs as the important drugs for autism treatment from the perspective of the topology structure of drug networks and drug interactions.

Table 5 shows the important drugs per clique in order of PageRank score. If there is a common drug between drugs in a clique and drugs interacting with existing drugs in the clique, we regarded them as the potentially valuable drug for autism. As a result, the following four drugs were identified: Tocilizumab (c15), Tacrolimus (c15), Prednisone (c2 and c3) and Sulfisoxazole (c23).

Table 5 34 cliques and the ranked first (or important drugs) identified from 1L1–1L5

Sequence of clique	*Important drugs*	*Sequence*	*Score*	*Interacting drugs*	*Sequence*	*Score*
c1	l-histidine	1	0.003006	quinine	1	0.004673
c10	nadh	1	0.003819	paclitaxel	1	0.004234
c11	pyridoxal phosphate	1	0.005194	cyclosporine	1	0.005538
c12	nadh	1	0.005765	acebutolol	1	0.025657
c13	nadh	1	0.006409	phenobarbital	1	0.007524
c14	fluconazole	1	0.003098	amiodarone	1	0.005481
c15	*tocilizumab*	2	0.004094	*tacrolimus*	2	0.006144
c15	*tacrolimus*	3	0.004094	*tocilizumab*	3	0.005954
c16	nadh	1	0.010165	mifepristone	1	0.008244

c17	fluconazole	1	0.003993	grepafloxacin	1	0.005522
c18	nadh	1	0.004649	azelastine	1	0.004595
c19	nadh	1	0.002921	cyclosporine	1	0.005398
c2	*prednisone*	4	0.00294	*prednisone*	9	0.005695
c20	nadh	1	0.001883	azelastine	1	0.006301
c21	xylometazoline	1	0.063041	quinine	1	0.007415
c22	nadh	1	0.00598	quinine	1	0.006526
c23	*sulfisoxazole*	7	0.0019	*sulfisoxazole*	4	0.006173
c24	nadh	1	0.007256	quinine	1	0.00642
c25	l-histidine	1	0.004622	quinine	1	0.003932
c26	fluconazole	1	0.003297	grepafloxacin	1	0.003214
c27	sulfamethizole	1	0.003141	quinine	1	0.005577
c28	cefacetrile	1	0.002241	haloperidol	1	0.005033
c29	fluconazole	1	0.004264	haloperidol	1	0.004575
c3	*prednisone*	2	0.003675	*prednisone*	5	0.003221
c30	neomycin	1	0.002252	azelastine	1	0.004905
c31	ofloxacin	1	0.002997	quinine	1	0.005361
c32	chlortetracycline	1	0.002816	azelastine	1	0.004518
c33	ciprofloxacin	1	0.00461	phenobarbital	1	0.003709
c34	fluconazole	1	0.004621	phenobarbital	1	0.003723
c4	l-histidine	1	0.004217	quinine	1	0.00293
c5	l-glutamic acid	2	0.008095	quinine	1	0.009528
c6	potassium iodide	2	0.007673	mifepristone	1	0.00837
c7	adenosine triphosphate	1	0.002106	paclitaxel	1	0.018128
c8	nadh	1	0.0023	mecamylamine	1	0.004701
c9	nadh	1	0.003239	paclitaxel	1	0.019765

Prednisone is a synthetic corticosteroid drug that is particularly effective as an immunosuppressant drug, but can cause depression or depressive symptoms in the short-term side effects, and cause Cushing's syndrome, steroid dementia syndrome, and depression in its long-term side effects. Tacrolimus is an immunosuppressive drug used mainly after allogeneic organ transplant to lower the risk of organ rejection. Recently, it has been used to treat segmental vitiligo in children, especially on face, and cause various neuropsychiatric problems such as loss of appetite, insomnia, posterior reversible encephalopathy syndrome, confusion, weakness, depression, vivid nightmares, cramps, neuropathy, seizures, tremors and catatonia in its side effects. However, although these two drugs are important, they may cause some symptoms of autism. Thus, when drug indications and adverse effect of drugs are considered, only Tocilizumab and Sulfisoxazole may have the potential to form a feasible hypothesis for autism treatment. Tocilizumab is "an immunosuppressive drug, mainly for the treatment of rheumatoid arthritis and systemic juvenile idiopathic arthritis, a severe form of arthritis in children" (Wikipedia, 2017). FDA approved Tocilizumab for the treatment of systemic juvenile idiopathic arthritis for children from age of two in 2011, and the

followed by European Medicines Agency in August 2011 (Aleman-Meza et al., 2005). Since disrupted communication between the nervous system and the immune system may cause various mental disorders, such as depression and autism(Kraneveld and Garssen, 2014), Tocilizumab has the potential to treat some symptoms of autism through immunotherapy. Sulfisoxazole is a sulfonamide antibacterial with an oxazole substituent. A short-acting sulfonamide antibacterial with activity against a wide range of gram- negative and gram-positive organisms, which is used to treat meningococcal meningitis (Drugbank, 2017b). As to Sulfisoxazole, Mecasermin, which is a drug "for the long-term treatment of growth failure in pediatric patients with Primary IGFD or with GH gene deletion", need to be mentioned (Drugbank, 2017a). Sulfisoxazole can improve Mecasermin's hypoglycemic activities to increase the therapeutic effect of the drug. Consequently, Sulfisoxazole may be the potentially important drug for autism medication.

Although those four drugs discovered by the proposed approach are likely to be a candidate treatment for Autism, it is a premature stage to argue that it requires the clinical trials for verification. Instead, we suggest a further study based on laboratory experiments to ensure suitability of those four drugs for Autism treatment. Out of the four drugs, Prednisone was mentioned as an effective treatment for Autism in a recent report (Massachusetts General Hospital, 2018). The report was based on an earlier indirect evidence in the literature regarding Prednisone and no published or on-going clinical trials were yet done in Autism treatment. Thus, it may be too early to determine whether Prednisone is the effective Autism treatment due to the risks involved compared to unknown benefits. However, it is certainly worthy of further study in laboratory experiments for validation. The report by Massachusetts General Hospital indirectly confirms the discovered drugs and the effectiveness of the proposed methods.

5. Conclusions

In the present study, we proposed a new approach to identifying drug associations and interaction among their therapeutic groups by topology-based network analysis and demonstrated how the proposed topology-driven trend analysis can be utilized for new drug discovery. We first analyzed the current trend of autism research by drug association and therapeutic group networks. In addition, we adopted various approaches to drug prediction for new autism treatment, which was based on the following two assumptions of literature-based prediction in this research: 1) Co-occurrence of entities has meaningful association. If drug (therapeutic group)-A co-occurs with drug (therapeutic group)-B, some kinds of relationship may exist between them, and the more frequently they co-occur, the closer they are related (Callon et al., 1991) and 2) An isolated drug (therapeutic group) in the network does not mean it has no relation with other drugs (therapeutic groups), while co-occurring pairs of drugs (therapeutic groups) are to be related.

Three major findings were reported in the present study. First, the topology-driven trend analysis by the pathfinding algorithm for associative drugs coupled with the clique percolation algorithm helped us extract the related therapeutic groups to detect cliques for new drug discovery throughout the different periods. In particular, by mapping drug associations to either 1st-level or 2nd-level ATC classification codes, the proposed topology-driven trend analysis uncovered the new therapeutic groups of drugs to recommend plausible new treatments for autism. Second, in grasping the overall research trend in autism research, the study found that use of drugs associated with cross-therapeutic groups became pervasive recently, and the proportion of multi-disciplinary research reached 70%

in the last 10 years. Specifically, the study reported on that nervous drugs are the most important to autism medication treatment. In addition, therapeutic groups like dermatological, the genito-urinary system and sex hormones were paid more attention by the autism research community. Drugs in the therapeutic groups such as nervous system, cardiovascular system, and dermatology have been steadily popular over last 10 years. Furthermore, a medication treatment structure consisting of "N," "C," "H," "S," "D," and "A" has been developed over last 10 years as well, and became the primary medication approach to treating autism. Third, we discovered that four drugs, Tocilizumab , Tacrolimus, Prednisone, and Sulfisoxazole, are the promising drugs for autism treatment, which is worthy of further study in laboratory experiments with formal assessment of possible effects on symptoms, and investigation of dosage effects and treatment duration for autism treatment.

The major limitation is that 13 drugs were not found in the ATC system, so we classified them to 2nd level according to their functions and components under the guidance of psychiatrists. Those drugs were marked by an asterisk (*) in Appendix 3.

The present study predicted four drugs for autism treatment based of the literature-based network. In the future, we will further verify the drugs discovered by literature reports, cell experiments, and laboratory experiments. We also plan to conduct content analysis to supplement the findings reported in the present study.

Acknowledgements

This research was supported by the Shanxi Provincial Education Department Research Program (2012221 and 2015BY36), China. This work was also partially supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT, and Future Planning through the National Research Foundation. Many thanks to the Web Science Lab team of Indiana University Bloomington for their discussion about this research. Many thanks to Qi Yu and Yujia Zhai, who strengthened my confidence and discussed the paper with me.

References

Adamo, S. A. (2008). 6–Bidirectional Connections Between the Immune System and the Nervous System in Insects. *Insect Immunology*, 129-149.

Aleman-Meza, B., Halaschek-Weiner, C., Arpinar, I. B., Ramakrishnan, C., & Sheth, A. P. (2005). Ranking complex relationships on the semantic web. *Internet Computing, IEEE, 9*(3), 37-44.

Association, A. P. (1994). Diagnostic and statistical manual. Washington, DC: American Psychiatric Association.

Baek, S.H., Lee, D., Kim, M., Lee, J.H., & Song, M. (2017). Enriching plausible new hypothesis generation in PubMed. PLoS ONE, 12(7), e0180539

Bauin, S., Michelet, B., Schweighoffer, M. G., & Vermeulin, P. (1991). Using bibliometrics in strategic analysis: "understanding chemical reactions" at the CNRS. *Scientometrics, 22*(1), 113-137.

Bianchi, L., Paganelli, F., Pettenati, M. C., Turchi, S., Ciofi, L., Iadanza, E., et al. (2014). Design of a RESTful Web information system for drug prescription and administration. *Biomedical and Health Informatics, IEEE Journal of, 18*(3), 885-895.

Björk, B.-C., Roos, A., & Lauri, M. Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options. In *ELPUB2008, 2008*

Bordons, M., Bravo, C., & Barrigón, S. (2004) Time-Tracking of the Research Profile of a Drug Using Bibliometric Tools, Journal of the American Society for Information Science and Technology, 55(5), 445–461.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155-205, doi:10.1007/bf02019280.Chang L., Yu J. X., and Qin L. (2013) Fast maximal cliques enumeration in sparse graphs. Algorithmica, 66(1):173–186.

Chen, B., Ding, Y., & Wild, D. J. (2012). Assessing drug target association using semantic linked data. *PLoS Comput Biol, 8*(7), e1002574.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery ☆. *Journal of Informetrics, 3*(3), 191-209.

Cheng T, Pan Y, Hao M, et al. (2014). PubChem applications in drug discovery: A bibliometric analysis. Drug Discov Today. 19:1751–1756. http://doi.org/10.1016/j.drudis.2014.08.008

Cohen, N. (1991). Bidirectional communication between the central nervous system (CNS) and the immune system. *Developmental & Comparative Immunology, 15*(3), 209.

Mecasermin (2017a). https://www.drugbank.ca/drugs/DB01277. Accessed 12/27 2017.

Drugbank (2017b). Sulfisoxazole. https://www.drugbank.ca/drugs/DB00263. Accessed 11/11 2017.

Evans, J. A. (2008). Electronic Publication and the Narrowing of Science and Scholarship. [10.1126/science.1150473]. *Science, 321*(5887), 395-399.

FDA (2009). Autism. http://www.fda.com/fdamd/autism.htm. Accessed 3/24 2016.

FDA (2012). National Drug Code Directory http://www.fda.gov/Drugs/default.htm. Accessed 12/22 2015.

Gordon, M. D., & Lindsay, R. K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the Association for Information Science & Technology, 47*(2), 116-128.

He, Q. (1999). Knowledge Discovery Through Co-Word Analysis. *Library Trends, 48*(1), 133-159.

health, B. i. f. i. a. s. M., & news, P. Brain inflammation found in autism. *Medicine/health Neurobiology Neurochemistry*.

Hong, Y., Yao, Q., Yang, Y., Feng, J., Wu, S., Ji, W., … Liu, Z. (2016). Knowledge structure and theme trends analysis on general practitioner research: A Co-word perspective. BMC Family Practice, 17, 10. http://doi.org/10.1186/s12875-016-0403-5

Jeong, Y. K., Heo, G. E., Kang, K. Y., Yoon, D. S., & Song, M. (2016). Trajectory analysis of drug-research trends in pancreatic cancer on PubMed and ClinicalTrials.gov. *Journal of Informetrics, 10*(1), 273-285, doi:http://dx.doi.org/10.1016/j.joi.2016.01.003.

Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258-263.

Jonsson, P. F., & Bates, P. A. (2011). Global topological features of cancer proteins in the human interactome. *Bioinformatics, 22*(18), 2291-2297.

Jonsson, P. F., Cavanna, T., Zicha, D., & Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *Bmc Bioinformatics, 7*(1), 1-13.

Kostoff, R. N. (2014). Literature-related discovery: common factors for Parkinson's Disease and Crohn's Disease. [journal article]. *Scientometrics, 100*(3), 623-657, doi:10.1007/s11192-014-1298-3.

Kraneveld, A. D., & Garssen, J. (2014). Targeting (Gut)-Immune-Brain Axis with Pharmaceutical and Nutritional Concepts: Relevance for Mental and Neurological Disorders. In G. Folkerts, & J. Garssen (Eds.), *Pharma-Nutrition: An Overview* (pp. 439-456). Cham: Springer International Publishing.

Larsen, P., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575-603.

Leydesdorff, L., D. Rotolo, D., Rafols, I. (2012). Bibliometric perspectives on medical innovation using the Medical Subject Headings (MeSH) of PubMed. Journal of the American Society for Information Science and Technology, 63:11, pp. 2239-2253

Lindsay, R. K., & Gordon, M. D. (1999). *Literature-based discovery by lexical statistics*: John Wiley & Sons, Inc.

Malone, R. P., Gratz, S. S., Delaney, M. A., & Hyman, S. B. (2005). Advances in drug treatments for children and adolescents with autism and other pervasive developmental disorders. *CNS drugs, 19*(11), 923-934.

Martin, A., Scahill, L., Klin, A., & Volkmar, F. R. (1999). Higher-functioning pervasive developmental disorders: rates and patterns of psychotropic drug use. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*(7), 923-931.

Massachusetts General Hospital. (2018). Lurie Center for Autism: Prednisone. https://www.massgeneral.org/children/services/lurie-center/Prednisone.aspx. Accessed 5/19 2018.

Modi, M. E., & Young, L. J. (2012). The oxytocin system in drug discovery for autism: animal models and novel therapeutic strategies. *Hormones and behavior, 61*(3), 340-350.

Nooy, W. d., Mrvar;, A., & Batagelj, V. (2005). Exploratory social network analysis with pajek. (pp. 334): Cambridge University Press.

Norén, N. (2011). Data-driven discovery: Case studies from patient records and spontaneous reports. https://www.pharmacoepi.org/pub/1c225a86-2354-d714-51af-5692f248677c. Accessed 5/18 2016.

Palla, G., Barabási, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature, 446*(7136), 664-667.

Rajpal DK, Kumar V, Agarwal P. (2011). Scientific literature mining for drug discovery: a case study on obesity. *Drug Development Research,72*(2):201-208.

Rajpal DK, Qu XA, Freudenberg JM, et al. (2014). Mining emerging biomedical literature for understanding disease associations in drug discovery. *Methods Mol Biol,1159*:171-206.

Smalheiser, N. R., & Swanson, D. R. (1996). Indomethacin and Alzheimer's disease. *Neurology, 46*(2), 583-583.

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics, 11*(1), 46-62, doi:https://doi.org/10.1016/j.joi.2016.11.001.

Song, M., Heo, G. E., & Ding, Y. (2015). SemPathFinder: Semantic path analysis for discovering publicly unknown knowledge. *Journal of Informetrics, 9*(4), 686-703, doi:https://doi.org/10.1016/j.joi.2015.06.004.

Stegmann, J., & Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. *Scientometrics, 56*(1), 111-135.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology & Medicine, 30*(1), 7.

Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the*

Association for Information Science & Technology,*38*(4):228-233.

Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology & Medicine*,*31*(4):526.

Swanson, D. R., & Smallheiser, N. R. (1999). Implicit Text Linkages between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery. *Library Trends, 48*(Summer), 48-59.

Toivonen, R., Onnela, J. P., Saramäki, J., Hyvönen, J., & Kaski, K. (2006). A model for social networks. *Physica A Statistical Mechanics & Its Applications, 371*(2), 851-860.

Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., et al. (2011). Finding complex biological relationships in recent PubMed articles using Bio-LDA. *Plos One, 6*(3), e17243.

Wei CP, Chen KA, Chen LC. (2014). Mining Biomedical Literature and Ontologies for Drug Repositioning Discovery. Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 373-384.

Wikipedia (2017). Tocilizumab. https://en.wikipedia.org/wiki/Tocilizumab2017. Accessed 11/11 2017.

Williams, R. S., Lotia, S., Holloway, A. K., & Pico, A. R. From Scientific Discovery to Cures: Bright Stars within a Galaxy. *Cell, 163*(1), 21-23, doi:10.1016/j.cell.2015.09.007.

Ying, D., Min, S., Jia, H., Qi, Y., Yan, E., Lin, L., et al. (2013). Entitymetrics: Measuring the Impact of Entities. *Plos One, 8*(8), 1-14.

Supplementary Material

Appendix 1. The number of synonyms per ATC Class. In this research, we used DrugBank and PharmGKB to expand synonyms and include drugs that interact with the 145 drugs selected for the study. The total number of synonyms is 6,624, and the number of synonyms per ATC is listed. Appendix 1 can be downloaded from http://informatics.yonsei.ac.kr/tsmm/download/Appendix1.docx.

Appendix 2. 34 cliques and important drugs identified from 1L1-1L5. Appendix 2 includes total 340 drugs and the top 10 ranked important drugs by PageRank per clique. The appendix can be downloaded from http://informatics.yonsei.ac.kr/tsmm/download/Appendix2.docx.

Appendix 3. Categories of the 145 drugs in the ATC system. All of the 145 drugs are classified into basic groups (1^{st} and 2^{nd} level) according to the ATC classification system, in which active substances were divided into different groups according to their therapeutic charactersitics. Appendix can be is downloaded from http://informatics.yonsei.ac.kr/tsmm/download/Appendix3.docx.