Understanding drug repurposing from the perspective of biomedical entities and their evolution: a bibliographic research using aspirin

Xin Li
Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

Justin F. Rousseau
Department of Population Health and Department of Neurology, Dell Medical School, The University of Texas at Austin, Austin, TX, U.S.A.

Ying Ding
School of Information, Dell Medical School, The University of Texas Austin, Austin, TX, U.S.A.

Min Song
Department of Library and Information Science, Yonsei University, Seoul, Korea

Wei Lu (corresponding author, Email: weilu@whu.edu.cn)
Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, Hubei, China
Abstract

**Background:** Nowadays, drug development is still a costly and time-consuming process with a low rate of success. Drug repurposing (DR) has attracted significant attention because of its significant advantages over traditional approaches, in terms of development time, cost, and safety. Entitymetrics, defined as bibliometric indicators based on biomedical entities (e.g., diseases, drugs and genes) studied in the biomedical texts, make it possible for researchers to measure the knowledge evolution and transfer of drug research.

**Objective:** The purpose of this study is to understand the drug repurposing from the perspective of biomedical entities (diseases, drugs and genes) and their evolution.

**Methods:** In the work reported in this paper, we extend the bibliometric indicators of biomedical entities mentioned in PubMed to detect potential patterns of biomedical entities in various phases of drug research, and investigate the factors driving DR. We use aspirin (acetylsalicylic acid) as the subject of the study, since it can be repurposed for many applications. We propose four easy, transparent measures based on entitymetrics: Popularity Index (P₁); Promising Index (P₂); Prestige Index (P₃); and Collaboration Index (CI), to investigate DR for aspirin.

**Results:** We find that the maxima of P₁, P₃ and CI are closely associated with the different repurposing phases of aspirin. These metrics enable us to observe the way in which biomedical entities interacted with the drug during the various phases of DR, and to analyze the potential driving factors of DR at the entity level. P₁ and CI are indicative of the dynamic trends of a specific biomedical entity over a long time period, while P₂ is more sensitive to immediate changes. P₃ reflects early signs of the practical value of biomedical entities, and could be valuable for tracking the research frontiers of a drug.

**Conclusions:** In-depth studies of side effects and mechanisms, fierce market competition, and advanced life science technologies are driving factors for DR. This study showcases the way in which researchers can examine the evolution of DR using entitymetrics, an approach which can be valuable for enhancing decision making in the field of drug discovery and development.

**Keywords:** drug repurposing; biomedical entities; entitymetrics; bibliometrics; aspirin; acetylsalicylic acid
Introduction

Background

Despite recent advances in life sciences and technology, drug development is still a costly and time-consuming process with a low rate of success [1]. Discovering a new drug usually takes more than 10 years and costs around $2 billion on average [2]. The number of targetable human genes is approximately 3,000, and the identification of serious and even deadly drug side effects is ongoing [3,4]. To overcome these difficulties, many researchers have turned to drug repurposing: the practice of identifying novel clinical indicators for existing marketed drugs [5–7].

The past decades have produced a few successful cases of drug repurposing. For examples, sildenafil, originally developed to treat cardiovascular disease, was unexpectedly discovered to be effective against erectile dysfunction (ED) [8]. Thalidomide, once used for morning sickness, has been repurposed for the treatment of multiple myeloma [9]; and metformin, originally a treatment for type II diabetes, has been studied for the treatment of depression, aging, obesity, and even cancer [10,11]. Beta blockers initially indicated for hypertension and topiramate originally used as an antiepileptic are both repurposed for migraineurs [12,13].

Because of its significant advantages over traditional approaches, in terms of development time, cost, and previous clinical studies, drug repurposing has attracted significant attention from pharmaceutical firms, scientists, and governments in recent years [7,14].

Methodologies for drug repurposing, and their successful applications have been widely discussed. Bin et al. [15] designed a system-based algorithm called Reverse Gene Expression Score (RGES) based on several large-scale publicly accessible datasets, and demonstrated the potency and efficacy of vorinostat, geldanamycin and gemcitabine for the treatment of liver cancers. Miao et al. [16] found that emricasan had an inhibitory effect on the Zika virus, by screening more than 6,000 compounds. With the rapid development of natural language processing (NLP) and deep learning techniques, robust solutions have been proposed recently and have demonstrated potential. Researchers have integrated more than 20 different datasets into a knowledge graph to predict potential drug and target pairs [17–19]. Hamilton et al. [20] queried drug-gene-drug interactions within a low-dimensional embedding of biomedical knowledge graphs to predict missing or unobserved links for drug repurposing. Chang et al. [21] proposed a novel deep learning model called
“CDRscan” that can successfully predict the feasibility of drug repurposing and recommend the most effective anticancer agents for an individual patient. Öztürk et al. [22] represented drugs and protein sequences using convolutional neural networks to predict the binding affinities of drug-target interactions.

Academic publications are produced at high volume, with around 3,000 new articles currently published per day [23]. No researcher or clinician can read and comprehend all of the relevant articles in their domain [24]. The “known” knowledge has turned into “unknown known” knowledge, with hidden information and patterns waiting to be discovered. This growing body of scholarly data opens a new era of exploiting literature and data to enable data-driven discovery [24]. Literature-based discovery, which connects disconnected entities in the PubMed literature, has been successful in identifying several cases of drug repurposing, such as fish oil for Raynaud’s syndrome, magnesium for migraine headaches, and proton pump inhibitors for atrial fibrillation [25–27]. Swanson [26] demonstrated that bibliometrics can be a useful approach to knowledge discovery, and recommended that his method could be extended to other disconnected sets of scientific literature to enable cross-disciplinary innovation [28]. With entitymetrics—bibliometric indicators based on entities studied in the medical literature—researchers without domain knowledge can understand the medical function of a drug [29], identify complex undiscovered biological relationships between drugs and targets [30], and detect implicit gene-gene relationship using the PubMed literature [31]. This research demonstrates the potential of applying bibliometrics to medicine to support data-driven discovery. It represents the next generation of bibliometric studies [32] and already shows great promise [33].

Objectives

In this research, we extend bibliometric indicators for biomedical entities mentioned in the PubMed literature to investigate drug repurposing. We use aspirin (salicylic acid) as the target drug. Aspirin is one of the most well-recognized and well-studied drugs, with a history dating back to 1,500 BC [34]. It was originally used as an analgesic to treat mild to moderate pain. It has been used clinically for the treatment of at least ten diseases, including coronary artery disease, cerebrovascular disease, peripheral arterial disease, preeclampsia, diabetes, colorectal cancer, Kawasaki disease, Alzheimer’s disease, and arthritis [34,35]. New indications for aspirin are still being reported [36–38]. Aspirin has a remarkably wide range of effects, and therefore provides an ideal case with which to study drug repurposing. The work described in this paper primarily aim to identify
patterns in the different repurposing phases of aspirin by analyzing the diseases, drugs and genes related to aspirin. We propose four measures based on entitymetrics: Popularity Index ($P_1$), Promising Index ($P_2$), Prestige Index ($P_3$) and Collaboration Index (CI), to identify the characteristics and patterns of DR for aspirin.

Related work

Drug repurposing

Drug repurposing has become a dynamic emerging field of drug discovery and development. According to Baker et al. [39], in 2018 nearly two-thirds of 35,000 drugs or compounds described in MEDLINE were investigated as potential treatments for diseases other than those for which they were originally indicated. Nearly 200 drugs have been investigated for repurposing, for more than 300 diseases. Many successfully repurposed drugs were discovered accidentally, such as the application of thalidomide to multiple myeloma [9], and sildenafil for erectile dysfunction (ED) [8].

Approaches have been proposed for the generation of hypotheses about novel drug-target interactions, and have been used to develop promising directions for subsequent validation of drug repurposing. In polypharmacology, researchers have proposed two types of hypotheses: (1) two drugs could be indicated for the same condition when they produce a similar gene expression profile, and (2) a disease could be one of the indications for a given drug when it has an opposite gene expression profile to that produced by the drug. The Connectivity Map (CMap), a database for more than 7,000 gene-expression profiles of 1,309 compounds, has been widely used in this context in previous work. Liu et al. [40] found that the anti-cancer drugs KM-00927 and BR-K75081836 can be used to inhibit histone deacetylase, using a systematic analysis tool, L1000FWD, and CMap. Kidnapillai et al. [41] used gene expression signature data and CMap to identify 10 drugs, including camptothecin, nimesulide, and recinnamine, which could be effective against bipolar disorder (BD).

In the field of genetics, association analysis has been extensively applied to the interactions between drug targets and diseases to increase the efficiency of drug repurposing. One of the most successful cases in the field of drug repurposing was based on a genome-wide association study (GWAS) [42]. Using GWAS-driven methods, Sanseau et al. concluded that 15.6 % of genes are the targets of marketed drugs. They found that
GWAS traits can be matched with the indications of drugs, and genes involved in pathogenesis have a high probability of being targets for drug repurposing [43]. Based on a strong association between the gene TNFSF11 and Crohn’s disease, the authors inferred, and subsequently confirmed, that dishubzumab, originally developed for the treatment of osteoporosis, can be used against Crohn’s disease [43]. Enrico and Pankaj [44] combined a CMap-based approach with perturbation of transcripational profiles and disease data from GWASs for target prioritization and drug repurposing. These researchers pointed out that the genetic evidence is important in maximizing the success rate of drug repurposing.

These methods in polypharmacology and genetics usually rely on the high-throughput screening of massive amounts of data related to compounds and targets. As knowledge about drug targets accumulates, and computational chemistry rapidly develops, simulations of the interactions between drugs and proteins have shown the potential to replace the traditional high-throughput screening. Dakshanamurthy et al. [45] proposed a proteochemometric method called TMFS to conduct molecular docking of over 3,000 FDA approved compounds across the crystal structures of more than 2,000 human targets. They found that mebendazole could be used for the inhibition of VEGFR2 kinase, and that celecoxib was a promising therapy for malignancies, because it binds an adhesion molecule, cadherin-11. Li et al. [46] designed a stand-alone approach to dock over 30 crystal structures of MAPK14 and BIM-8 with all drugs from DrugBank, and found that nilotinib, as a potential inhibitor of MAPK14, could be a cure for inflammatory diseases.

Another significant source of drug repurposing is drug side effects. Typical instances of side effect-based drug repurposing include the use of sildenafil for erectile dysfunction [8] and the application of exenatide acetate to obesity [47], both of which were “happy accidents”. Recently, Lun and Pankaj [48] generated human phenotypic profiles for drugs, based on over 3,000 side-effect relationships extracted from PharmGKB, and employed naïve Bayes methods to identify new indications for drugs according to their side effects. This study also suggested that the use of side effects is a type of clinical phenotypic assay, and side effects should be rationally investigated to predict repurposing opportunities for drugs. Ye et al. [49] contend that drugs with similar side effects could share the same indications, because they may have the same or similar mechanisms of action. Using a side-effect similarity-based drug-drug network, they transformed drug repurposing into an
information retrieval issue and successfully obtained the top five indications of 1,234 drugs approved by the FDA.

With the rise of machine learning and deep learning in computer science and bioinformatics, the problem of drug repurposing has been addressed using approaches such as classification [50,51], link prediction [52,53], entity prediction [52], and path prediction [18,54]. Liang et al. [52] represented biomedical entities and their relationships in a heterogeneous network using graph2vec and knowledge2vec [55], and employed a cascade learning model to find potential interactions between drugs, genes, diseases, and treatments. They found that vitamin D could be a treatment for prostate cancer. Fu et al. [54] treated drug repurposing as a binary classification problem, and combined the meta-path-based topological features of biomedical entities in Chem2Bio2RDF and a supervised machine learning model to predict links between drugs and targets. They found that the intrinsic feature selection Random Forest algorithm can be valuable for selecting significant topological features for the prediction of links between drugs and genes.

**Big scholarly data for medical knowledge discovery**

Traditionally, knowledge discovery in medical domains has relied on first-hand observation such as epidemiological statistics, follow-ups and laboratory-generated experimental data [24]. A large number of research papers are published daily, posing significant challenges for scientists wishing to have a comprehensive understanding of their domain [24]. The “known” knowledge has turned into “undiscovered public knowledge”, with patterns and information waiting to be uncovered. This large body of literature and data also provides rich opportunities for researchers to undertake data-driven knowledge discovery. The usefulness of literature-based discovery has been demonstrated in many previous research projects. For instance, the “ABC” model proposed by Swanson in 1986 was used to discover relationships between biomedical entities, such as Raynaud’s syndrome and fish oil [25], migraine headaches and magnesium [26], and fibrillation and proton pump inhibitors [27]. The “ABC” model is co-occurrence-based and is based upon the premise that seemingly unrelated concepts A and C could be related when there is a concept B related to both A and C [27]. Since Swanson’s research, various modifications of the “ABC” model have been proposed to discover hidden relationships among biomedical concepts in PubMed, such as ontology-based entity mapping [56], network-based entity extraction [57] and semantic path-based storytelling [58]. The “ABC”
model and its variants indicate that bibliometrics can be a valuable method for medical knowledge discovery in the era of big scholarly data.

Knowledge graphs of big scholarly data can contain nodes representing biomedical entities such as diseases, drugs, genes, pathways, and cell lines, and non-biomedical entities such as authors, institutions, articles, journals, conferences, and keywords. Edges in the graph can represent the relationships between the biomedical entities in the literature. Lv et al. [59] established a therapeutic knowledge graph for autism, using drug entities and MeSH terms extracted from about 20,000 articles relating to autism, published between 1946 and 2015. They proposed a novel topology-driven method incorporating various graph-analytical techniques for drug discovery, and concluded that Tocilizumab, Sulfisoxazole, Tacrolimus and Prednisone were promising for the treatment of autism. Ding et al. [29] constructed an entity-entity citation graph to highlight the significance of biomedical entities embedded in literature for future knowledge discovery. Researchers have also integrated big scholarly data with other publicly accessible biomedical datasets, such as DrugBank [60], Gene Ontology [61], and SIDER [62], to form a comprehensive knowledge graph for medical knowledge discovery. A typical example is the Chem2Bio2RDF database, created by integrating more than 20 chemogenomic datasets with PubMed. Wang et al. [30] proposed a novel algorithm called Bio-LDA to automatically extract latent topics in life sciences, and identified relationships and patterns among compounds, genes and diseases from Chem2Bio2RDF. He et al. [63] designed a graph mining algorithm to predict potential relationships between different biomedical entities. The case they studied demonstrated that the anti-diabetic drug Rosiglitazone has cardiovascular-related side effects.

Entitymetrics, an entity-driven bibliometric method, and the next generation of citation analysis [29,32], make it possible for researchers without domain knowledge to measure the impact, usage, and transfer of knowledge entities embedded in academic texts for further knowledge discovery [32]. Ding et al. [29] built an entity-entity citation graph based on articles related to metformin, and detected most of the known interactions of metformin with biomedical entities. Williams et al. [64] recognized and quantified relationships between academic discoveries and major advances in the domain of two new drugs, ipilimumab and ivacaftor, to enhance government support and public understanding. Zhu et al. [65] established paper-entity, entity-entity co-occurrence and entity-specific networks based on the scientific literature to investigate the evolution of
hepatic carcinoma at a fine-grained level. Lv et al. [59] discovered new indications for drugs using topology-driven trend analysis of drug-drug and drug-indication networks. These studies demonstrate the potential of the application of bibliometric methods to data-driven discovery in medical domains.

Drug repurposing, as one of the most significant issues in the field of medical knowledge discovery, has been extensively investigated in [17,23,24,27,28,54–56, 63]. In this research, we extend the bibliometric indicators for biomedical entities described in the PubMed literature to understand the process of drug repurposing.

Methods

Data Collection

Papers on aspirin-related research published between 1951 and 2018 were collected from PubMed. Since aspirin is known by many names, the search terms were chosen from DrugBank, RxNorm and MeSH terms [33, 60]. The final search query is shown in Box 1. Non-journal articles, non-English articles, letters and editorial commentaries were excluded. In total, 63,387 publications from PubMed were downloaded in XML format.

Box 1 Search query used for retrieving aspirin-related publications

(((aspirin) OR ( acetylsalicylic acid) OR (acid, acetylsalicylic) OR (“2-(acetyloxy)benzoic aci”) OR (acylpyrin) OR (aloixprimum) OR (colfarit) OR (dispril) OR (easprin) OR (ecotrin) OR (endosprin) OR (magneacyl) OR (micristin) OR (polopiri) OR (polopiryn) OR (solupr) OR (soluprs) OR (zorprin) OR (acetsal) OR (2-acetoybenzenecarboxylic acid) OR (2-acetoxybenzoic acid) OR (acetylsalicylate) OR (acetylsalicylsäure) OR (“acide 2-(acétyloxy)benzoïqu”) OR (acide acétylsalicylique) OR (ácido acetilsalicílico) OR (acidum acetylsalicylicum) OR (aspirina) OR (azetylsalizylsäure) OR (o-acetooxybenzoic acid) OR (o-acetylsalicylic acid) OR (o-carboxyphenyl acetate) OR (salicylic acid acetate) ) AND ("1951"[PDAT] : "2018"[PDAT]))

To better understand the drug repurposing process of aspirin, the relevant research was divided into four phases based on previous studies [34,35] and expert suggestions: (1) 1951-1960: the original use; (2) 1961-1990: in-depth studies of pharmacological mechanisms and side effects; (3) 1991-2000: repurposing for
cardiovascular diseases; and (4) 2001-2018: repurposing for other diseases, such as colorectal cancer and breast cancer. These phases can also be observed from the evolution and trends of the publications, as shown in Figure 1 and Table 1.

Before extracting biomedical entities, all articles were parsed to obtain PMIDs, publication years, titles, abstracts, authors, journals, and institutions, using a dom4j XML parser written in Java. Then we used SpaCy to do the preprocessing (such as removing the punctuation and stop words) of titles and abstracts in the NLP pipeline. In addition, a novel and reliable method of author name disambiguation proposed by Lerchenmuller and Sorenson [66] was used to count distinct authors.

<table>
<thead>
<tr>
<th>Phases</th>
<th>Time span</th>
<th>Number of publications</th>
<th>Number of authors</th>
<th>Avg. number of authors</th>
<th>Number of journals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. original use</td>
<td>1951-1955</td>
<td>208</td>
<td>318</td>
<td>1.76</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>1956-1960</td>
<td>299</td>
<td>498</td>
<td>1.88</td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>1951-1960</td>
<td>507</td>
<td>794</td>
<td>1.83</td>
<td>218</td>
</tr>
<tr>
<td>2. in-depth studies of</td>
<td>1961-1965</td>
<td>748</td>
<td>1310</td>
<td>2.01</td>
<td>301</td>
</tr>
<tr>
<td>pharmacological mechanisms and</td>
<td>1966-1970</td>
<td>1268</td>
<td>2167</td>
<td>2.12</td>
<td>418</td>
</tr>
<tr>
<td>side effects</td>
<td>1971-1975</td>
<td>2766</td>
<td>4880</td>
<td>2.40</td>
<td>696</td>
</tr>
<tr>
<td></td>
<td>1976-1980</td>
<td>3797</td>
<td>7419</td>
<td>2.71</td>
<td>895</td>
</tr>
<tr>
<td></td>
<td>1981-1985</td>
<td>4395</td>
<td>10011</td>
<td>3.16</td>
<td>1033</td>
</tr>
<tr>
<td></td>
<td>1986-1990</td>
<td>4470</td>
<td>11600</td>
<td>3.50</td>
<td>1101</td>
</tr>
<tr>
<td></td>
<td>1961-1990</td>
<td>17444</td>
<td>31787</td>
<td>2.90</td>
<td>2153</td>
</tr>
<tr>
<td>3. repurposing for cardiovascular diseases</td>
<td>1991-1995</td>
<td>5164</td>
<td>14044</td>
<td>3.69</td>
<td>1256</td>
</tr>
<tr>
<td></td>
<td>1996-2000</td>
<td>6353</td>
<td>17694</td>
<td>4.10</td>
<td>1314</td>
</tr>
<tr>
<td></td>
<td>1991-2000</td>
<td>11517</td>
<td>28818</td>
<td>3.91</td>
<td>1798</td>
</tr>
<tr>
<td>4. repurposing for other</td>
<td>2001-2005</td>
<td>8099</td>
<td>27784</td>
<td>4.22</td>
<td>1719</td>
</tr>
<tr>
<td>diseases</td>
<td>2006-2010</td>
<td>9366</td>
<td>35313</td>
<td>4.94</td>
<td>1974</td>
</tr>
<tr>
<td></td>
<td>2011-2015</td>
<td>10436</td>
<td>44603</td>
<td>5.78</td>
<td>2410</td>
</tr>
<tr>
<td></td>
<td>2016-2018</td>
<td>6018</td>
<td>30796</td>
<td>6.73</td>
<td>1881</td>
</tr>
<tr>
<td></td>
<td>2001-2018</td>
<td>33919</td>
<td>118857</td>
<td>5.33</td>
<td>3865</td>
</tr>
<tr>
<td>Total</td>
<td>63387</td>
<td>171559</td>
<td>4.39</td>
<td>5443</td>
<td></td>
</tr>
</tbody>
</table>

**Biomedical entity Extraction**

The biomedical entity extraction module provided by the Biomedical Entity Search Tool (BEST) [67], a dictionary-based biomedical information extraction tool based on sophisticated information retrieval approaches, was deployed to extract entities such as diseases, drugs, and genes. The dictionary of BEST is built from 12 different public sources, including NCBI Entrez Gene, DrugBank, T3DB, Animal TFDB,
Therapeutic Target DataBase, PubChem, and MeSH [67]. We obtained 1,472 unique disease names, 1,640 unique drug names and 3,184 unique gene names from the titles and abstracts. Table 2 shows the top 10 biomedical entities of three different types and their frequency of appearance in PubMed articles.

Table 2 Top 10 biomedical entities in aspirin-related publications during 1951-2018.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Diseases</th>
<th>Frequency</th>
<th>Drugs</th>
<th>Frequency</th>
<th>Genes</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>coronary disease</td>
<td>2707</td>
<td>clopidogrel</td>
<td>6223</td>
<td>COX-2</td>
<td>3957</td>
</tr>
<tr>
<td>2</td>
<td>asthma</td>
<td>2277</td>
<td>ticlopidine</td>
<td>5433</td>
<td>CD143</td>
<td>1495</td>
</tr>
<tr>
<td>3</td>
<td>diabetes</td>
<td>1840</td>
<td>heparin</td>
<td>4391</td>
<td>COX-1</td>
<td>1179</td>
</tr>
<tr>
<td>4</td>
<td>hypersensitivities, drug</td>
<td>1342</td>
<td>indomethacin</td>
<td>3462</td>
<td>Plasminogen</td>
<td>1131</td>
</tr>
<tr>
<td>5</td>
<td>ulcer, gastric</td>
<td>1146</td>
<td>warfarin</td>
<td>3457</td>
<td>LDLCQ3</td>
<td>1081</td>
</tr>
<tr>
<td>6</td>
<td>cerebral ischemia</td>
<td>1135</td>
<td>vitamin f</td>
<td>2760</td>
<td>LPLA2</td>
<td>1047</td>
</tr>
<tr>
<td>7</td>
<td>intracranial vascular disorder</td>
<td>1133</td>
<td>dipyridamole</td>
<td>2232</td>
<td>GPIIib</td>
<td>1017</td>
</tr>
<tr>
<td>8</td>
<td>ischemic heart disease</td>
<td>1090</td>
<td>adenosine</td>
<td>2188</td>
<td>P2Y12</td>
<td>855</td>
</tr>
<tr>
<td>9</td>
<td>carcinomas, colorectal</td>
<td>1085</td>
<td>acetaminophen</td>
<td>2099</td>
<td>tPA</td>
<td>748</td>
</tr>
<tr>
<td>10</td>
<td>rheumatoid arthritis</td>
<td>832</td>
<td>prostacyclin</td>
<td>1498</td>
<td>TNF-α</td>
<td>629</td>
</tr>
</tbody>
</table>

**Entitymetric indicators for biomedical entities (P3C)**

In order to quantify and visualize the academic importance of individual biomedical entities, four transparent and easy entitymetric indexes (P3C) are developed: the Popularity Index, the Promising Index, the Prestige Index, and the Collaboration Index. These indicators can be considered as the extensions of the indicators proposed by Kissin for measuring the academic interest of a drug or a technique at the article level [33, 68]. In this paper, we adapt them from the perspective of biomedical entities with the goal to understand drug repurposing. Different from Kissin’s indicators, our indicators not only focus on the articles on a given drug, but also consider the changes in indicators of biomedical entities (e.g., diseases, drugs, and genes) and non-biomedical entities (e.g., authors) that are related to the given drug. Detailed explanations of these measures are as follows:

- **Popularity Index** ($P_1$) of a certain biomedical entity reflects the percentage of publications discussing that biomedical entity among all publications in a research field during a specific period, usually five years. The popularity of a biomedical entity $i$, *Popularity Index* ($i$), is given by:

$$P_1(i) = \frac{N_i}{N_T} \times 100\% \quad (1)$$
where \( N_i \) is the number of publications relating to an entity \( i \) in a period and \( N_T \) represents the total number of publications in the research field during the same period. An increase in \( P_1 \) indicates growing academic interest in \( i \) in the field.

- **Promising Index (\( P_2 \))** of a biomedical entity is the change in the popularity of an entity \( i \) in a research field between two continuous periods. The promising index of a specific biomedical entity \( i \), \( P_2(i) \), is expressed as:

\[
P_2(i) = \frac{N_i}{N_T} - \frac{N_{pi}}{N_{pT}}
\]  

where \( \frac{N_{pi}}{N_{pT}} \) refers to the popularity of the entity \( i \) in the research field during a previous period of the same length as \( N_i \). \( P_2 \) reflects the change in the academic interest in a biomedical entity in a research field in two time periods. When \( P_2(i) > 0 \), it means the academic interest in \( i \) increases, and *vice versa*.

- **Prestige Index (\( P_3 \))** is defined as the ratio of the number of publications about a specific biomedical entity published in the top journals compared to the number of publications about the same entity in all journals that were indexed by PubMed during the same time period. The prestige of a biomedical entity \( i \), \( P_3(i) \), is calculated as:

\[
P_3(i) = \frac{N_{H20}}{N_i} \times 100\%
\]  

where \( N_{H20} \) represents the number of publications on \( i \) in the top 20 journals during the same period as \( N_i \). In this study, the top 20 journals were selected based on the journal impact factor (JIF) and specialty areas. These journals can be divided into two categories: multidisciplinary journals and specialty journals. Fourteen multidisciplinary journals, including *JAMA*, *The Lancet*, *BMJ*, and similar publications, are common for all diseases, drugs, and genes that were studied in this paper. The other six journals are highly associated with aspirin-related specialty areas, such as *Circulation*, *Blood*, and *The European Heart Journal*. The full list of the top 20 journals is shown in Table S1. The Prestige Index (\( P_3 \)) reflects the potential significance of a specific biomedical entity. Continuing high
prestige scores could be an early signal of the success of entity-related drug discovery or repurposing [68]. We employed a threshold of 5% to indicate that P3 was of interest [68].

- **Collaboration Index (CI)** of a biomedical entity reflects the percentage of the number of distinct authors of articles discussing this entity out of all of the distinct authors in the research domain over a period of time. The collaboration index of a biomedical entity \( i \), \( CI(i) \), is calculated by:

\[
CI(i) = \frac{N_{Ai}}{N_{AT}} \times 100\%
\]  

(4)

where \( N_{Ai} \) is the number of distinct authors of the publications relating to \( i \) in a period, and \( N_{AT} \) represents the total number of distinct authors in the field in the same period. The CI reflects the research strength of entity \( i \) in a research field, and a threshold of 5% indicates a level of interest [68].

**Results**

**Overview of aspirin-related studies**

![Figure 1](image)

Figure 1 Number of aspirin-related studies in PubMed over time. The background colors indicate the four phases of aspirin research.

Figure 1 shows an overview of aspirin-related research in PubMed from 1951 to 2018. The red and blue lines represent the percentage and absolute numbers of articles in PubMed per year, respectively. The details of
publications, authors, and journals are shown in Table 1. In the course of the evolution of aspirin, Phase 1 (1951-1960) produced 507 articles, most of which were published in journals covering pharmacy-related or general medicine-related topics (Table 1 and Figure S1 in Supplementary Information). Research in Phase I focused on the anti-inflammatory and antipyretic uses of aspirin, and this phase marks the original use of aspirin.

In Phase 2 (1961-1990), a turning point can be identified in 1967, after which the number of relevant papers per year grew dramatically until 1986. Several significant pharmacological discoveries related to aspirin occurred during this period, including the anti-platelet effect [69], the mechanism of inhibition of prostaglandin synthesis [70], and the acetylation of the cyclo-oxygenase enzyme [71]. The percentage of aspirin-related articles in PubMed reached its peak in 1981, at about 0.32%, and then decreased. Kune et al. (1988) reported that aspirin could effectively reduce the incidence of colorectal cancer [72], after which the percentage began to rise again. After 1975, articles began to occur frequently in journals covering specialty areas, such as Circulation and Thrombosis Research. We identify this phase as the investigation of the in-depth studies of pharmacological mechanisms and side effects of aspirin.

Phase 3 (1991-2000) witnesses a steady and stable growth in the number and percentage of aspirin-related articles per year in PubMed (Figure 1). Compared to the first ten years (1951-1960), the number of articles increased by over 22 times, and the number of distinct authors increased by over 36 times. As shown in Figure S1, in both 1991-1995 and 1996-2000, four of the top five journals were cardiovascular-related journals. We thus identify this phase as repurposing for cardiovascular diseases.

In Phase 4 (2001-2018), the number of articles per year grew continuously and reached its peak (2,164) in 2015, but the percentage significantly reduced (Figure 1). From Table 1, we note that the numbers of articles, distinct authors and journals were all higher than those of the previous three periods. The average number of authors in this period had exceeded the total average (4.39). Journals covering other topics, for examples, Cancer Management and Research, Drugs & Aging, and World Neurosurgery, were increasingly represented (Figure S1), demonstrating that aspirin had been experimentally applied to many other diseases. We thus mark this phase as repurposing for other diseases.
To analyze drug repurposing through all four phases from the biomedical entity perspective, we first compute the P3C indicators of the top 10 diseases, drugs and genes in the cohort of aspirin articles during the period 1951-2018. The results show that there are distinct patterns of these indicators in different repurposing phases. To describe these patterns in detail, we reorganize the thirty biomedical entities (the top 10 disease, drugs, and genes) into the four phases of aspirin research, according to when each achieved its maximum Popularity Index, which indicates the focus of research in the field of aspirin. In each phase, we further analyzed the change patterns of P3C indicators for the most popular biomedical entities, to investigate the features of different phases of drug repurposing, the association between entities and P3C indicators, and the possible factors driving drug repurposing at the biomedical entity level.

**Before repurposing**

Only “rheumatoid arthritis” (RA) reached its maximum Popularity Index ($P_1$) in Phase 1, at 9.36%, as shown in Figure 2 (1), and then exhibited a downhill trend for the rest of the three phases, and reaching a low of 0.63% in 2016-2018. As shown in Figure 2 (2), for the Promising Index ($P_2$) of RA, there is only one significant rise of more than 0 in all four phases: 0.06 in 1951-1955 (Phase 1). This observation indicates that the popularity of RA in 1951-1955 increased by 6% compared to that of 1945-1950. It can also be observed from Figure 2 (3) that the Prestige Index ($P_3$) of RA was more than 5% during 1951-1980, and reached its maximum in Phase 1 (25, 1960-1965), indicating that one quarter of the papers studying RA were published in the top 20 journals in the aspirin domain in Phase 1. In the next three phases, the $P_3$ peaked twice, in Phase 2 (1971-1975) and Phase 3 (2001-2005), possibly relating to the discovery of the mechanism of anti-inflammatory and RA-induced cardiovascular diseases. Similar to $P_1$, as shown in Figure 2 (4), the Collaboration Index (CI) of RA peaked in 1956-1960 (40.44%), then declined to 1.02% in 2016-2018, indicating that around 40% of authors in Phase 1 were studying RA, but only about 1.02% authors still worked on the same disease in Phase 4.
In summary, in Phase 1, the P1, P2, P3 and CI of RA reached their maxima, or showed a significant rise, indicating that RA was the disease upon which most research was focused in the aspirin domain at this time. However, the value of these indicators showed profound declines in the next three phases, which means that aspirin was studied in relation to other diseases, and is thus an ideal example of drug repurposing.

The scientific basis for repurposing
Figure 3 The Popularity Index ($P_1$) of the biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the four phases of aspirin research.

As shown in Figure 3, there are nine top biomedical entities in the aspirin domain which reached their maximum $P_1$ in Phase 2, including three diseases (“asthma,” “hypersensitivities, drug” and “ulcer, gastric”) and six drugs (indomethacin, acetaminophen, dipyridamole, vitamin F, adenosine and prostacyclin”. The three diseases can all be side effects of aspirin, while the six drugs can be divided into three categories: (1) competitors of aspirin, that is, indomethacin and acetaminophen, which are analgesic and antipyretic drugs respectively, with fewer side effects; (2) the antiplatelet drug dipyridamole; and (3) precursor substances in the pathway of the mechanism of action of aspirin: vitamin F, adenosine and prostacyclin. In contrast with RA, the $P_1$ of these biomedical entities increased from Phase 1, peaked in Phase 2, and then decreased, indicating that
the side effects and the mechanisms of aspirin were studied in depth in Phase 2. The P₁ of indomethacin in 1976-1980 (16.75) was the highest among these nine entities in Phase 2, and vitamin F in 1981-1985 (11.19) ranked second.

Figure 4 shows the P₂ of these nine biomedical entities in the aspirin domain over time. The P₂ of three side effects had a significant rise of more than zero in Phase 2, 1961-1965 and 1976-1980 for “asthma”, 1961-1965 for “hypersensitivities, drug”, and 1961-1965 for “ulcer, gastric”, indicating that interest in the side effects of aspirin increased sharply. The time periods in which the P₂ of the six drugs showed significant rises are generally later than those of the side effects, such as 1971-1975 for indomethacin and 1981-1985 for prostacyclin. This observation indicates that the discovery and in-depth studies of side effects may have
positive effects on the discovery of the mechanism of action of aspirin, as well as the development of its alternatives with fewer side effects.

Figure 5 Prestige Index (P3) of biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the four phases of aspirin research.

Figure 5 shows the P3 of these nine biomedical entities in the aspirin domain, demonstrating a feature common to all nine entities: a gradual decline with fluctuation in P3 after reaching a maximum in Phase 1 or Phase 2. The P3 of “hypersensitivities, drug” and “ulcer, gastric” had a highest initial value in Phase 1, revealing that both side effects had been taken seriously by researchers in Phase 1. The P3 of “hypersensitivities, drug” in 1956-1960 (33.33%) was higher than that of RA in 1956-1960 (25.00%). In 2011-2015, the P3 of only two entities are over the 5% threshold: 5.82% for adenosine and 10% for prostacyclin. In the aspirin domain, papers studying these two entities published in the top 20 journals comprised more than
5% of papers published in all of the journals indexed by the PubMed in 2011-2015. This observation indicates that the two entities were still important foci of research in the aspirin domain.

Table 3 Intervals between the time periods of the maxima of $P_1$ and $P_3$. 

<table>
<thead>
<tr>
<th>Biomedical entity</th>
<th>Time period of the maximum of $P_1$ (T1)</th>
<th>Time period of the maximum of $P_3$ (T2)</th>
<th>T1-T2</th>
</tr>
</thead>
</table>

It can be observed from Figure 3, Figure 5 and Table 3 that $P_3$ on average achieved their maxima 10.7 years earlier than $P_1$. In particular, for “hypersensitivities, drug” and “ulcer, gastric”, the intervals can be as long as 20 years. This observation indicates that the Prestige Index can reflect an early sign of academic interest into biomedical entities, a phenomenon which could be potentially valuable for tracking the research frontiers of a drug.

The results of the CI of these nine biomedical entities in the aspirin domain are presented in Figure 6, which shows that the CIs for these biomedical entities have similar trends to $P_1$ over time. Indomethacin achieved the highest maximum of CI in 1976-1980 (19.79%) among all nine biomedical entities during 1951-2018, indicating that it became a strong competitor to aspirin as an analgesic agent in Phase 2. This result also demonstrates that during the last five-year period (2011-2015), only two of the nine entities’ CI were over 5%, meaning that the two entities were still the subject of research of a considerable number of scientists (more than 2,230) in the aspirin research community in 2011-2015. The two biomedical entities include “asthma” (6.21%) and adenosine (5.50%).

Based on the observation of P3C in Phase 2 and previous studies on aspirin [34,35], we can conclude that on one hand, the in-depth investigation of side effects and the mechanism of action of aspirin provided the knowledge basis and research direction for drug repurposing. On the other hand, due to the market competition
from other drugs, as well as the serious side effects, pharmaceutical companies attempted to discover new indicators for aspirin, in order to maintain the sales volume of aspirin.

Figure 6 Collaboration Index (CI) of biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the four phases of aspirin research.

**Repurposing aspirin for cardiovascular-related diseases**

In Phase 3, five top biomedical entities comprising four diseases and one drug reached their maximum Popularity Index, as shown in the first row of Figure 7. The four diseases were all cardiovascular related, including “coronary disease (CD)” (P1 = 18.88% in 1996-2000), “cerebral ischemia” (P1 = 2.57% in 1996-2000), “intracranial vascular disorder” (P1 = 5.73% in 1991-1995) and “ischemic heart disease (IHD)” (P1 = 3.01% in 1996-2000). Compared with Figure 2 and Figure 3, the P1 of the previous ten biomedical entities that
peaked in the first or second phases were considerably lower than that of CD, indicating that the cardiovascular-related disease was the focus of the aspirin domain in that time. CD is often referred as IHD, and is the most common cardiovascular-related disease worldwide; similarly, “cerebral ischemia” and “intracranial vascular disorder” represent the same condition, commonly known as stroke. These conditions were reported to be the first and second most common causes of death worldwide in the early 21st century [73]. The demand for the prevention and treatment of such fatal diseases could be one of the factors that drive the repurposing of aspirin for cardiovascular-related diseases.

Figure 7 The P3C of biomedical entities on the cardiovascular diseases in aspirin domain over time. The background colors show the four phases of aspirin research

The only drug that reached its maximum P1 in Phase 3 is heparin (11.92% in 1996-2000). As one of the most common anticoagulant drugs, heparin has always been the reference drug for repurposing aspirin to treat cardiovascular-related diseases, which could be the reason for the increase in the academic interest in heparin
in the aspirin domain. There was another peak of heparin in Phase 2 (5.03%, 1971-1975), which could be related to an increase in research into the mechanisms of the anti-platelet effect of aspirin in Phase 2.

The second row of Figure 7 shows the changes in P₂ of these five biomedical entities over time. All five biomedical entities demonstrated a significant rise in Phase 3. “Coronary disease” and “cerebral ischemia” had risen in 1991-1995, and “intracranial vascular disorder”, “ischemic heart disease” and “heparin” rose in 1991-1995. The P₂ of the two entities also showed significant increases in Phase 2: 0.02 in 1976-1980 for “coronary disease”, and 0.10 in 1971-1975 for “heparin”, consistent with the fact that aspirin was clinically used for coronary disease before the discovery of its anti-platelet effect.

The pattern of P₃ for these five entities over time is displayed in the third row of Figure 7. All five biomedical entities reached their maxima in Phase 2, earlier than the maximum of P₁. “Coronary disease” reached a maximum in 1971-1975, and heparin in 1961-1965. The difference from the previous phases is that the P₃ of these five biomedical entities peaked again in Phase 3. For instance, “coronary disease” peaked in 1991-1995, and heparin in 1991-1995, indicating that these biomedical entities were important topics of research in both Phase 1 and Phase 3.

The last row of Figure 7 shows the CI of five biomedical entities during 1951-2018, in which the CI demonstrated a dynamic trajectory very similar to that of the Popularity Index. The maximum of “coronary disease” in Phase 3 is highest at 22.91% in 1996-2000, indicating that “coronary disease” attracted the greatest share of the authors in the aspirin domain. “Coronary disease” and “cerebral ischemia” in Phase 4, and heparin in Phase 2 and 4 surpassed the threshold value of 5%. The CI of “cerebral ischemia” steadily grew after Phase 3, showing a different trend from the other four biomedical entities, which increased in the Phase 1 and Phase 2, peaked in Phase 3, and then dramatically decreased. This observation may illustrate that “cerebral ischemia” unlike the other biomedical entities is still increasing in its popularity and collaboration so there is still more increases expected to come.

**Repurposing aspirin for other diseases**

In Figure 8, there are fifteen biomedical entities which reached their maximum Popularity Index in Phase 4. Unlike the previous phases, most of the biomedical entities were genes, and can be divided into three categories according to the diseases to which they are related: (1) inflammatory-related genes: COX-2,
LPLA2, and TNF-α; (2) cardiovascular-related genes, including COX-1, CD143, Plasminogen, LDLCQ3, GPIIb, P2Y12 and tPA; and (3) cancer-related genes such as TNF-α, COX-2, COX-1 and LPLA2. These observations indicate that aspirin was actively studied for these three aspects of diseases from the perspective of genes in Phase 4. In particular, the maximum P1 of COX-2 was the highest among these fifteen biomedical entities at 21.97% in 2001-2005, revealing that COX-2 was considered to be very important in the aspirin domain in that time.

Figure 8 also shows that the P1 of two diseases peaked in Phase 4. One is “diabetes”, whose P1 in 2006-2010 was 6.83%. In fact, as early as 1875, Ebstein and Muller discovered that aspirin had the effect of lowering blood sugar. Inspired by this observation, scientists have since been trying to use aspirin for the treatment of diabetes [74]. There are several peaks in the P1 of “diabetes” in previous phases. In the 21st century it has been recommended that patients with diabetes who have an increased risk of cardiovascular disease take aspirin as a primary preventative [5,75]; this could be the reason why the academic interest in “diabetes” in the aspirin domain increased again. The other disease is “carcinomas, colorectal”. Its P1 peaked in 2001-2005, and then increased significantly after a small decline in 2006-2010, a pattern which is very different from other diseases in the aspirin domain. Repurposing aspirin for the treatment of colorectal carcinomas appears to be a focus of research in the aspirin domain today. The P1 of three drugs also peaked in Phase 4, including the anti-platelet drugs clopidogrel and ticlopidine, which are competitors of aspirin as anti-platelet drugs [35]; and warfarin, which is an anticoagulation drug that is similar to heparin and has been found to be superior to aspirin for secondary prevention of ischemic stroke with nonvalvular atrial fibrillation [76,77].
Figure 8 Popularity Index ($P_1$) of biomedical entities on repurposing aspirin for other diseases over time. The background colors show the four phases of aspirin research.

Figure 9 Promising Index ($P_2$) of biomedical entities on repurposing aspirin for other diseases over time. The background colors show the four phases of aspirin research.
Figure 10 Prestige Index ($P_3$) of biomedical entities on repurposing aspirin for other diseases over time. The background colors show the four phases of aspirin research.

Figure 9 presents the changes in $P_2$ of these fifteen biomedical entities over time. All of the genes demonstrate a rise of more than 0 in Phase 4. Unlike these genes, the diseases and drugs showed several significant rises of over 0 in different phases, which reflects a longer history of research in the aspirin domain. For example, 1956-1960, 1996-2000 and 2001-2005 for “diabetes”, 1996-2000, 2001-2005 and 2006-2010 for clopidogrel, and 1971-1975 and 1991-1995 for warfarin.

The changes in $P_3$ of these fifteen biomedical entities over time are shown in Figure 10, from which we can make two observations. First, the $P_3$ of these biomedical entities demonstrated that the time period of the maximum of $P_3$ was much earlier than that of the maximum of $P_1$. Second, unlike the biomedical entities noted in previous sections, the diseases and drugs had two or more significant peaks in different phases. For instance, “diabetes” had peaks of 42.86% in 1956-1960, 25.00% in 1971-1975 and 14.22 in 1996-2000; and “carcinomas, colorectal” had peaks of 33.33 in 1981-1985, 15.91 in 1991-1995, and 14.15 in 2006-2010. These numbers indicate that these entities attracted considerable interest in the field of aspirin research, and
high impact papers on these conditions were published. However, the genes usually had only one peak in $P_3$ in the third or fourth phases, illustrating that these genes are relatively new topics in the aspirin domain.

The CI data for these fifteen biomedical entities are presented in Figure 11, which shows that the maximum of COX-2’s CI is the highest, at 34.37%, in 2001-2015 for COX-2, denoting that COX-2 was the focus of aspirin research in Phase 4; the research and development of Vioxx, a selective COX-2 inhibitors with less side effects, may be one of the reasons [78]. The CI of two drugs, clopidogrel (25.54% in 2001-2015) and ticlopidine (20.74 in 2006-2010), reveals fierce competition between aspirin and these alternative anti-platelet drugs. This competition could have driven the repurposing of aspirin for other diseases, especially cancers, which have an urgent demand for effective treatment.

**Discussion**

**Principal Findings**

This study examines drug repurposing from the perspective of the evolution of biomedical entities, using aspirin as the study subject. It is of paramount importance for drug discovery to identify the factors that drive
repurposing, as well as to identify potential patterns among biomedical entities in various phases of the drug research timeline. The main contribution of this paper is twofold. First, this paper proposes four entitymetric indices (P3C) to quantify changes in academic interest in biomedical entities and to reveal the fine-grained process of drug repurposing. Second, we divide aspirin research into four phases, including the original use (1951-1960), in-depth studies of pharmacological mechanisms and side effects (1961-1990), repurposing for cardiovascular-related diseases (1991-2000) and repurposing for other diseases (2001-2018), taking into consideration three fine-grained perspectives—disease, drug, and gene—which contribute to a comprehensive understanding of the features of the repurposing process.

Our entitymetric results indicate that aspirin is representative of the process of drug repurposing. The research findings can be summarized as follows: in Phase 1, aspirin was routinely used to ease pain, fever and inflammation, and was often used in the treatment of rheumatoid arthritis [34], with a P3C which peaked in 1951-1960. Despite the widespread use of aspirin, at this stage its mechanism of action was not well understood [34]. In Phase 2, the side effects and mechanisms of actions of aspirin were studied extensively, as shown by the maxima of P₁ and CI, as well as a significant rise of P₂ for the relevant biomedical entities in 1961-1990. The anti-platelet effect [69], inhibition of prostaglandin synthesis [70] and the acetylation effect on the enzyme cyclo-oxygenase [71] were uncovered. These discoveries had laid a solid knowledge foundation for the successful repurposing of aspirin. The highest P₁ in 1961-1990 was for indomethacin (16.75), denoting a fierce competition of aspirin for its original use. This could be one of the factors contribute to the repurposing of aspirin.

In Phase 3, aspirin was successfully used for several cardiovascular-related diseases because of its anti-platelet effect [79]. The related diseases and drugs achieved their highest values of P₁ and CI, as well as significant rises in P₂ in 1991-2000. As these diseases are the most common diseases worldwide, according to WHO statistics [73], the demand for the prevention and treatment of the fatal diseases is also probably another factor driving drug repurposing. In the last phase, there were a large number of studies suggesting the use of aspirin for other diseases, especially colorectal cancer [36]. The greatest difference from previous phases is that aspirin was studied at the level of genes. Ten genes had a maximum P₁ and CI, as well as an apparent rise in P₂ in 2001-2018. This observation could indicate that the development of modern science and technology,
such as gene sequencing, molecular simulation and deep learning, accelerates the process of drug repurposing of aspirin. Meanwhile, two fatal diseases—diabetes and colorectal carcinoma—as well as three competitive drugs of aspirin as an anti-platelet agent—clopidogrel, ticlopidine and warfarin (an anticoagulant and competed with aspirin for stroke prevention)—also had peak $P_1$ and CI values, as well as a great rise in $P_2$.

Methodologically, in this study we develop four entitymetrics, and demonstrate how to use them to investigate the process of drug repurposing. The results demonstrate that the maximums of $P_1$, $P_3$ and CI are closely associated with the different phases of research into aspirin repurposing. The $P_1$ and CI metrics can indicate dynamic trends in academic interest in a given biomedical entity over a long time period. For instance, long-lasting rises in $P_1$ and CI signal interest in repurposing, while $P_2$ is more sensitive to immediate changes in academic interest in a specific biomedical entity, since it takes into consideration data from the two most recent periods. Moreover, the $P_3$ can reflect a research focus far earlier than the other three indices, which means that a continuing high $P_3$ may be valuable as an early signal of the emergence and transfer of research topics in drug research. If the Prestige Index does indeed have predictive power, it could be due to the involvement of top domain experts in the peer review of manuscripts in top journals with high impact factors [80,81]. Additionally, due to their easy implementation and interpretability, these indices can be applied in multiple domains, such as drug assessment, drug discovery and pharmacovigilance.

Limitations and Future Directions

There are several limitations in the current paper. First, the data included in our analysis are limited to articles indexed in PubMed. Some real-world data, such as EHR, clinical trials, and social media, in which aspirin and its related biomedical entities were mentioned, should be included. In our future work, we will use different types of data sources for studying drug repurposing and take into account other entities related to drugs, including other biomedical entities, such as pathways, proteins, and cells, and non-biomedical entities, such as authors, institutions, and countries. The landscape of collaborations between academic and pharmaceutical could affect the drug repurposing process. Second, there are several ways of measuring the impact of a journal, such as Impact Factor (IF) and Relative Citation Ratio (RCR). Third, this paper mainly focuses on investigating the repurposing journey of aspirin, but we didn’t test whether it can be used for predicting future drug repurposing. In the future studies, we will evaluate different impact measures of a journal and choose a
proper one fitting better to the chosen drug. Furthermore, we also aim to test the proposed metrics on the other drugs to understanding their repurposing journeys (e.g., metformin) to see whether there exist generalized patterns in different repurposing processes.

Acknowledgements

This study was supported by the Major Project of the National Social Science Foundation of China (17&ZDA292). The support provided by China Scholarship Council (CSC) during a visit of Xin Li to Indiana University Bloomington (No. 201806270047) is acknowledged. This work was also partly supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT, and Future Planning through the National Research Foundation. The authors are also grateful to the anonymous referees and editors for their invaluable and insightful comments.

Conflicts of Interest

None declared.

References


24. Ding Y, Stirling K. Data-driven Discovery: A new era of exploiting the literature and data. J Data Inf Sci


42. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 Jun;447(7145):661–678. [doi: 10.1038/nature05911]


47. Leinung MC, Grasso P. [d-Leu-4]-OB3, a synthetic peptide amide with leptin-like activity, augments the effects of orally delivered exenatide and pramlintide acetate on energy balance and glycemic control in insulin-resistant male C57BLK/6-db/db mice. Regul Pept 2012 Nov 10;179(1–3):33–38. PMID:22960403


Supplementary Information

Figure S1 Changes in the number of journals with aspirin-related publications during 1951-2018. The top five journals and their frequencies are indicated using heatmaps for every five-year period.
<table>
<thead>
<tr>
<th>NO.</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>THE NEW ENGLAND JOURNAL OF MEDICINE</td>
</tr>
<tr>
<td>2</td>
<td>LANCET</td>
</tr>
<tr>
<td>3</td>
<td>JAMA</td>
</tr>
<tr>
<td>4</td>
<td>BRITISH MEDICINE JOURNAL</td>
</tr>
<tr>
<td>5</td>
<td>NATURE</td>
</tr>
<tr>
<td>6</td>
<td>NATURE REVIEW. DRUG DISCOVERY</td>
</tr>
<tr>
<td>7</td>
<td>NATURE MEDICINE</td>
</tr>
<tr>
<td>8</td>
<td>SCIENCE</td>
</tr>
<tr>
<td>9</td>
<td>PHARMACOLOGICAL REVIEWS</td>
</tr>
<tr>
<td>10</td>
<td>ANNALS OF INTERNAL MEDICINE</td>
</tr>
<tr>
<td>11</td>
<td>THE JOURNAL OF CLINICAL INVESTIGATION</td>
</tr>
<tr>
<td>12</td>
<td>PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCE OF THE UNITED STATES OF AMERICA</td>
</tr>
<tr>
<td>13</td>
<td>TRENDS IN THE PHARMACOLOGICAL SCIENCES</td>
</tr>
<tr>
<td>14</td>
<td>THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS.</td>
</tr>
<tr>
<td>15</td>
<td>CIRCULATION</td>
</tr>
<tr>
<td>16</td>
<td>JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY</td>
</tr>
<tr>
<td>17</td>
<td>BRITISH JOURNAL OF HAEMATOLOGY</td>
</tr>
<tr>
<td>18</td>
<td>BLOOD</td>
</tr>
<tr>
<td>19</td>
<td>EUROPEAN HEART JOURNAL</td>
</tr>
<tr>
<td>20</td>
<td>JOURNAL OF THROMBOSIS AND HAEMOSTASIS</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Entity Name</td>
<td>$P_1$ (%)</td>
</tr>
<tr>
<td>coronary disease</td>
<td>(71-90)</td>
</tr>
<tr>
<td>diabetes</td>
<td>2.34 (56-60)</td>
</tr>
<tr>
<td>ulcer, gastric</td>
<td>33.33 (56-60)</td>
</tr>
<tr>
<td>cerebral ischemia</td>
<td>(76-80)</td>
</tr>
<tr>
<td>rheumatoid arthritis</td>
<td>9.36 (56-60)</td>
</tr>
<tr>
<td>heparin</td>
<td>25 (61-65)</td>
</tr>
<tr>
<td>warfarin</td>
<td>44.44 (61-65)</td>
</tr>
<tr>
<td>vitamin f</td>
<td>15.87 (71-75)</td>
</tr>
<tr>
<td>dipyriramole</td>
<td>9.42 (81-85)</td>
</tr>
<tr>
<td>acetaminophen</td>
<td>6.07 (81-85)</td>
</tr>
<tr>
<td>prostacyclin</td>
<td>8.42 (81-85)</td>
</tr>
</tbody>
</table>
Table S2 summarized the peaks of the P1, P3 and CI as well as the rise of P2 for all the top 30 biomedical entities. The details can be found in the supplementary information section, including the phase 2 (1961-1990, the scientific basis for repurposing), the phase 3 (1991-2000, repurposing aspirin for cardiovascular-related diseases) and phase 4 (2001-2018, repurposing aspirin for other diseases).

<table>
<thead>
<tr>
<th>Entity</th>
<th>(81-85)</th>
<th>(76-80)</th>
<th>(86-90)</th>
<th>(96-00)</th>
<th>(01-05)</th>
<th>(11-15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COX-2</td>
<td></td>
<td></td>
<td></td>
<td>34.58</td>
<td>21.97</td>
<td>34.37</td>
</tr>
<tr>
<td>CD143</td>
<td></td>
<td></td>
<td></td>
<td>10.17</td>
<td>5.68</td>
<td>10.36</td>
</tr>
<tr>
<td>COX-1</td>
<td></td>
<td></td>
<td></td>
<td>14.63</td>
<td>3.86</td>
<td>6.88</td>
</tr>
<tr>
<td>Plasminogen</td>
<td>(86-90)</td>
<td>(91-95)</td>
<td></td>
<td>4.45</td>
<td>17.16</td>
<td>9.79</td>
</tr>
<tr>
<td>LDLCQ3</td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td>3.93</td>
<td>8.38</td>
</tr>
<tr>
<td>LPLA2</td>
<td></td>
<td></td>
<td></td>
<td>3.97</td>
<td>17.41</td>
<td>8.27</td>
</tr>
<tr>
<td>GPIIb</td>
<td>53.33</td>
<td></td>
<td></td>
<td>25</td>
<td>9.72</td>
<td>4.59</td>
</tr>
<tr>
<td>P2Y12</td>
<td></td>
<td></td>
<td></td>
<td>23.07</td>
<td>3.84</td>
<td>8.17</td>
</tr>
<tr>
<td>tPA</td>
<td>33.33</td>
<td></td>
<td></td>
<td>3.33</td>
<td>2.68</td>
<td>7.36</td>
</tr>
<tr>
<td>TNF-α</td>
<td>14.28</td>
<td></td>
<td></td>
<td>1.44</td>
<td>9.84</td>
<td>7.27</td>
</tr>
</tbody>
</table>

Note: The numbers in parentheses indicate the years.