Using Entity Metrics to Understand Drug Repurposing

Xin Li¹², Ying Ding¹², Wei Lu¹

¹School of Information Management, Wuhan University, Wuhan, Hubei, China; ²School of informatics and computing, Bloomington, Indiana, USA

Abstract

Understanding the process of drug repurposing is critically significant for drug development. In this paper, we employ extracted bio-entities to detect the features of different phases in drug repurposing. We proposed a transparent and easy entitymetric indicator for bio-entities, i.e., Popularity Index, to quantify and visualize the dynamic changes in academic interests of bio-entities. By taking aspirin as an example, the results display specific profiles of drug repurposing and the evolution of bio-entities in the different phases of drug research, which would potentially be valuable for pharmaceutical companies and scholars to successfully discover and repurpose drugs.

Introduction

Despite advances in life science and technology, drug development is still a costly and time-consuming process with low rates of success¹. Discovering a new drug usually takes over 10 years and around $2 billion². Meanwhile, the number of targetable human genes is reported to be limited at approximately 3,000³, whereas that of serious and deadly drug side-effects increases⁴.

To overcome these difficulties, drug repurposing, defined as the practice of identifying novel clinical indications for existing marketed drugs, is often employed⁵,⁶. The past decades have witnessed a few successful cases of drug repurposing, for examples, sildenafil originally developed to treat cardiovascular diseases was unexpectedly discovered to be an efficacious cure for erectile dysfunction (ED)⁷; and thalidomide, once used for morning sickness, was repurposed for multiple myeloma⁸. In recent years, the methodologies for drug repurposing and corresponding cases of success have been widely discussed in studies⁵,⁶,⁹,¹⁰, however, few studies have offered an in-depth understanding of the process of drug repurposing.

In this study, taking aspirin as a case study, we provide an in-depth understanding of drug repurposing from the perspective of bio-entities and their evolutions. For one thing, as the minimum domain knowledge carriers, diseases entities can provide clear and in-depth understanding of knowledge evolution and communication in specific cases. For another, bio-entities are timelier and more accurately to capture the dynamic changes of academic interest from the quantity perspective or track the semantic differences of research topics from the content dimension.

In particular, we develop a transparent and easy entitymetric indicator (i.e., Popularity Index) for bio-entities, to quantify and visualize the dynamic evolution of academic interests in aspirin-related disease entities in the different phases. Then, we quantitatively detect the patterns of these bio-entities interacted with aspirin and how they co-occurred, to understand the roles that these entities played in the process of the drug repurposing of aspirin.

The reminder of this paper is organized as follows. First, the “Related work” reviews the relating studies, giving attention to the drug repurposing and entitymetrics. Then, “Methods” details the data and the approaches used in our paper. Next, “Results” presents the results of this study. Finally, “Discussion and conclusions” concludes with our findings and the limitations and proposes several directions for future works.

Related work

Drug repurposing

Drug repurposing has attracted growing academic attention in recent years, however, most of the related studies are about the approaches to how to discovery or predict repurposable drug candidates, such as literature-based discovery¹⁰,¹¹, network analysis¹⁴,¹⁵, and machine learning⁵,¹⁰. Few studies have investigated the specific process of drug repurposing from a fine-grained perspective, as well as the characteristics in its different phases. From a bibliometrics perspective, Baker et al. (2018) reviewed drug repurposing from a macro level, by analyzing over 25 million articles in PubMed, and concluded that > 60% drugs have been used for > 1 disease, including 189 drugs have been tried as a cure for > 300 diseases¹⁶. Pushpakom et al. (2019) detected the progress and challenges of drug repurposing using system review, and summarized that methods employed in drug repurposing can be divided into computational and experimental methods⁵. However, these methods can’t reveal the repurposing process of a specific drug from its original use to repurposed indications, from the perspective of research interest evolution, or the roles played by the
different entities in these phases. Hence, the aim of this paper is to understand the drug repurposing process by analyzing bio-entities and their evolution.

**Entitymetrics**

Entitymetrics, originally proposed by Ding et al. (2013), refers to measure the impact, usage, and transfer of knowledge entities embedded in academic texts for further knowledge discovery. It is regarded as an entity-driven bibliometric method as well as the next generation of citation analysis. Researches focusing on using entitymetrics to discovery knowledge and its evolution of a biomedical domain include those of Ding et al. (2013), who established an entity-entity citation graph based on the articles in the metformin domain and detected most of the interactions of metformin with bio-entities for drug discovery. Williams et al (2015) recognized and quantified relationships between academic discoveries and major advances in the domain of two new drugs (ipilimumab and ivacaftor), to enhance the government support and public understanding for life science studies. Zhu, Min and Yan (2016) established the paper-entity, entity-entity co-occurrence and entity-specific networks based on scientific literatures, to identify the evolution of hepatic carcinoma at a fine-grained level. Lv et al. (2019) discovered new indications for drugs using topology-driven trend analysis of the drug-drug and drug-indication networks.

This study differs from the previous works in two respects. On one hand, the aim of our study is to detect the process of drug repurposing from the bio-entity perspective by tracing the transfer and usage of concrete knowledge entities. On the other hand, we design an easy entitymetric index to quantify and visualize the dynamic changes of the academic interest of bio-entities, which provides a more accurate measurement to understand the knowledge flows in drug repurposing.

**Methods**

**Data source and bio-entities extraction**

Papers on aspirin-related research were collected from PubMed for the period from 1951 to 2018. The search strategy was based on a set of search terms chosen by checking the DrugBank and MeSH terms, discussions within all the authors, and suggestions from domain experts. We also excluded the articles that were non-journal articles, non-English articles, letters and editorial commentaries. In this way, we obtained 63,387 publications from PubMed in XML format. Then, all articles were parsed to obtain PMIDs, publication years, titles, and abstracts, using a developed dom4j XML parser written in Java. The results were stored in a local relational database for bio-entity extraction and further analysis.

We used spaCy (https://spacy.io/) to preprocess the unstructured texts such as titles and abstracts. Then we employed the biomedical entity extraction module provided by Biomedical Entity Search Tool (BEST), a dictionary-based biomedical information extraction tool, to extract disease entities. Finally, we obtained 1472 unique disease names. Table 1 shows the top 10 diseases and their document frequency.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Diseases</th>
<th>Frequency</th>
<th>Rank</th>
<th>Diseases</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>coronary disease</td>
<td>2707</td>
<td>6</td>
<td>cerebral ischemia</td>
<td>1135</td>
</tr>
<tr>
<td>2</td>
<td>asthmas</td>
<td>2277</td>
<td>7</td>
<td>intracranial vascular disorder</td>
<td>1133</td>
</tr>
<tr>
<td>3</td>
<td>diabetes mellitus</td>
<td>1840</td>
<td>8</td>
<td>ischemic heart disease</td>
<td>1090</td>
</tr>
<tr>
<td>4</td>
<td>hypersensitivities, drug</td>
<td>1342</td>
<td>9</td>
<td>carcinomas, colorectal</td>
<td>1085</td>
</tr>
<tr>
<td>5</td>
<td>ulcer, gastric</td>
<td>1146</td>
<td>10</td>
<td>rheumatoid arthritis</td>
<td>832</td>
</tr>
</tbody>
</table>

**Entitymetric indicator (Popularity Index)**

Popularity Index (PI) of a certain bio-entity means the percentage of publications discussing it among all publications in a research field during a period (usually 5 years). The popularity of a bio-entity $i$, *Popularity Index* ($\text{PI}(i)$), is given by:

$$\text{PI}(i) = \frac{N_i}{N_T} \times 100\%$$
where $N_i$ is the number of publications relating to $i$ in a period, $N_T$ represents the total number of publications in the research field during the same period. If the $P_i$ increases, it illustrates that the academic interest of $i$ is growing in the field.

**Results**

**Overview of aspirin-related studies**

Figure 1. The overview of aspirin-related studies in PubMed.

Figure 1 shows the overview of aspirin-related researches in PubMed from 1951 to 2018, in which the red and blue lines represent the percentage and absolute numbers of articles, respectively. We can observe that the development of aspirin research can be roughly divided into four phases: 1951-1960, 1961-1990, 1991-2000, and 2001-2018, according to the evolution and content of publications. In 1951-1960, most of the 507 articles were published in journals covering pharmacy-related or general medicine-related topics, and the research focused on the anti-inflammatory and antipyretic use of aspirin, therefore marking this phase as the original use of aspirin.

In 1961-1990, a turning point can be identified in 1967, from which the number of related papers per year had been drastically grown until 1986. This may be due to the discovery of significant pharmacological actions of aspirin, including the anti-platelet effect (O’Brien, 1967), the mechanism of inhibitory on prostaglandin synthesis (Vane, 1971), and the acetylation effect on cyclo-oxygenase enzyme (Roth, 1975). The percentage of aspirin-related articles in PubMed reached its peak in 1981 (about 0.32%) and then it decreased. Kune et al. (1988) reported that aspirin could effectively reduce the incidence of colorectal cancer, from which the percentage began to rise again. Thus, we called this phase as the in-depth studies of pharmacological mechanisms and side effects of aspirin.

The third phase witnessed a steady and stable growth in the number and percentage of aspirin-related articles per year in PubMed. Compared to 1951-1960, the number of the articles and the distinct authors increased by over 22 times and 36 times, respectively. Meanwhile, four-fifths of the top five high-frequent journals were cardiovascular-related journals. We thus call this phase as the repurposing for cardiovascular diseases.

In the final phase, we note that the numbers of articles were over the sum of the previous three periods. Journals covering other topics, for example, CANCER MANAGEMENT AND RESEARCH, DRUGS & AGING and WORLD NEUROSURGERY, had attracted increasing attention, demonstrating that aspirin had been tried repurposing for many other diseases. We thus mark this phase as the repurposing for other diseases.

**Repurposing from the perspective of diseases**

Figure 2 represents the article-related Popularity Index (PI) of the top 10 diseases in the aspirin domain for the past 68 years, from which we can find that in 1951-2018 the evolution of disease entities shows a flow pattern, that is, from inflammatory-related diseases and side effects to cardiovascular diseases, then to other diseases, especially colorectal carcinoma.
Figure 2. The Changes of the Popularity Index (PI) of the top 10 diseases in the aspirin-related studies during the period of 1951-2018.

Specifically, in 1951-1960, aspirin was mainly used for “rheumatoid arthritis”\textsuperscript{26}, whose PI reached its highest values (9.36). In 1961-1990, with the discoveries of the pharmacological actions and the deepening understanding of the side effects of aspirin\textsuperscript{22-24}, the PI of side effects, i.e., “asthmas” (4.68), “hypersensitivities, drug” (4.65) and “ulcer, gastric” (3.73), arrived their peaks. At the same time, the anti-palate effect of aspirin was proven in this period\textsuperscript{23} and attracted much attention in the aspirin domain, the PI of cardiovascular diseases including “coronary disease”, “cerebral ischemia”, “intracranial vascular disorder” and “ischemic heart disease”, started growing fast. In 1991-2000, aspirin has been widely used for cardiovascular diseases\textsuperscript{27}, and their PI reached their peaks in that period: 11.06 for “coronary disease”, 2.57 for “cerebral ischemia”, 5.73 for “intracranial vascular disorder” and 3.01 for “ischemic heart disease”. In 2001-2018, aspirin was repurposed for many other diseases, the PI of all diseases were monotonically decreasing except for “carcinomas, colorectal”, indicating that colorectal cancer is the research focus of aspirin domain in that period.

In particular, as early as 1875, aspirin was found to have hypoglycemic effects and soon after was used for diabetes\textsuperscript{28}. Then, with the in-depth studies of diabetes, patients with diabetes was proven to have an increased risk of coronary disease, thus they were suggested to take aspirin for the primary prevention in the 20s\textsuperscript{29}. Therefore, multiple peaks of the PI of “diabetes” (2.34 in 1956-1960, 2.12 in 1981-1985 and 6.83 in 2006-2010) can be observed in the Figure 2.

Discussion and conclusion

This is the first research that employed a specific entitymetric indicator for bio-entities to investigate the process of drug repurposing at a fine-grained level. It demonstrated that the Popularity Index used in this paper can successfully reflect the evolution of the indications of aspirin through the appearance of disease entities in academic articles. The results displayed here illustrate that over the past 68 years aspirin research has experienced four phases, i.e., the original use (1951-1960), in-depth studies of pharmacological mechanisms and side effects (1961-1990), repurposing for cardiovascular diseases (1991-2000) and repurposing for other diseases (2001-2018).

There are several limitations to the current paper that need to be considered. Firstly, the findings on drug repurposing presented in this study may not be generally accepted, and thus a more comprehensive investigation of drug repurposing from other drugs (e.g., “metformin”, “chloroquine”, and “sildenafil”) is required. Nevertheless, we believe that the observations in this paper can provide scientists a motive for exploring the drug repurposing paradigm in specific areas, which can form reasonable future directions to study. Secondly, the data included in our analysis are restricted to PubMed articles, however, more real-world data (RWD), such as EHR, clinical trial databases, FDA files,
and even social media, in which “aspirin” and its related biomedical entities were stated, should be also analyzed in the future. Third, this paper only focused on the disease entities, but did not take into account other entities related to drugs, including other biomedical entities (e.g., drug, gene, pathway, and protein) and non-biomedical entities (e.g., authors, institutions, and countries). Therefore, we believe there is still much room for further studies, and we expect interesting findings in the follow-up works.

References