A Study on Drug Repositioning Clue Extraction Pipeline Based on Network Analysis in Social Network Services

Wongyun Lee¹, Seunghee Lee², Jong-Yeup Kim³, Suehyun Lee⁴

¹Researcher, Department of Healthcare Data Science Center, Konkuk University Hospital, Daejeon; ²Research Professor, Department of Healthcare Data Science Center, Konkuk University Hospital, Daejeon; ³Professor, Department of Healthcare Data Science Center, Konkuk University Hospital, Daejeon; ⁴Professor, Department of Computer Engineering, College of IT Convergence, Gachon University, Seongnam, Korea

Introduction

In this New drug development requires a period of at least 10 years and a lot of investment, but the probability of success is extremely low. However, drug repositioning takes relatively little time and money to apply drugs that have been approved for use to other diseases [1]. Thalidomide can be mentioned as a representative example of drug repositioning. Thalidomide was developed for the treatment of morning sickness in pregnant women, but its use was banned due to the side effect of giving birth to deformed babies. However, it was found to be efficacious in multiple myeloma and was relaunched on the market with a new utility [2].

On the other hand, there are cases in which drug reaction have been used as indications for other conditions and have progressed to drug repositioning. As a typical example, aspirin was found to be effective in preventing cardiovascular disease in addition to fever and labor, and it was recreated as an anti-thrombotic drug, and Minoxidil’s hair growth side effect was repositioned as a treatment for hair loss [3].

In this study, we utilize data from social network services to explore drug reactions for drug repositioning. Recently, attempts are being made to repurpose drugs using social media. In fact, there have also been studies evaluating the feasibility of using patient reviews on social media to identify potential candidates for drug repositioning [4,5]. We propose a methodology that can promote social network analysis, search for keywords of drug reaction, and establish a basis for drug repositioning (Figure 1).

Methods

A process was established to conduct this research. Each step of selecting a target drug for analysis, collecting data, cleaning the collected data, and analyzing the cleaned data was performed in sequence.
Data collecting

Naver, the Korea’s largest portal site, was used to collect data on drug users’ reactions. Twitter has the advantage of collecting drug response data because it is simple and less burdensome to express users’ thoughts than other platforms from the user’s point of view. In fact, there have been cases of collecting data on drug reactions using Twitter [6], but it is not suitable to collect data on commercially available drugs because the user ratio of Twitter is not high in Korea. The selection of target drugs was selected as common drugs that are easy for users to recognize, drugs with high disease prevalence and prescriptions, in order to ensure a sufficient amount of data.

Search for selected drugs among Naver Cafe’s posts by keyword, and collect comment data on the category, URL, content of posts, and content of all posts inquired. The collection used Python’s Selenium module. Selenium is a module for crawling web pages and extracting only the data you need [7]. In the case of Naver Cafe, only cafe members are given access to data, but during crawling, data can be accessed without authority, so data can be collected smoothly.

Data preprocessing

Tokenization

In the case of Korean, due to the characteristics of mixed words, there is a feature that words are attached with surveys and used differently [8], so we proceeded with tokenization based on morphemes. We used the Korean morphological analyzer Mecab from the KoNLPy library, a Python package for Korean natural language processing [9]. At this time, the words for all reactions to drugs were registered in the morphological analyzer as a Mecab user dictionary, and the texts included in the user dictionary were tagged with proper nouns during morphological analysis. The user dictionary was built using the WHO-ART ver.092 Korean version codebook distributed by the Korea Agency for Drug Safety. WHO-ART is an international classification system for adverse drug reaction terminology, used for ADR (Adverse Drug Reaction) reporting, where it is easy to pre-construct all terms for drug reactions [10]. For example, depending on drug use, there may be two reactions: “deteriorating diabetes” and “relieving diabetes,” which can be seen as a side effect of the drug, and “relieving diabetes” can be seen as the effect of the drug. In this study, since the positive negative expression for keywords was not considered, and keywords such as “diabetes” were focused on excluding positive negative, it was judged that keywords of side effects terms could be used as reactions to drugs. After that, we extracted the words that exist in the WHO-ART ver.092 codebook from the tokenized text. For the extracted words, BoW (Bag of Words), which is a frequency-based word expression method, was used to quantify the number of times each word appeared, and a matrix of DTM (Document-Term Matrix) was constructed.

Filter defining

In order to reduce the keywords that can be used for drug repositioning, a dictionary has been established to define keywords that are insufficient. The significance of this study on drug repositioning is to establish the basis for drug repositioning by capturing informal symptoms reported through users, although they are not officially known symptoms from specific drugs. Therefore, using data from the Korea Pharmaceutical Information Service (KPIS), three dictionaries were established to distinguish official symptoms of drugs that cannot be used in this study.

1) Drug efficacy: Since the known efficacy of a drug is the most frequently detected keyword, it can be seen that it is the furthest from the purpose of this study to detect unknown symptoms. To pre-
vent this error, a drug efficacy filter dictionary was constructed with data from pharmacy sources.

(2) Caution symptom: Symptoms that users should check before taking the drug are not related to the drug. It was determined that the keyword could interfere with the purpose of this study to observe the reaction of a drug, and a filter dictionary for drug use was established based on the data from the Drug Information Service.

(3) Drug side effect: We tried to improve the quality of the available keywords by removing known side effects in advance according to drug use.

a) (over 1% of clinical trials): Among the side effects of the selected drugs, it was judged that the keywords for unknown symptoms could be applied to those closer to repositioning of the drug. A filter dictionary was constructed using drug side effect information (adverse side effect keywords that showed a response of 1% or more in clinical trials) from the Pharmaceutical Information Institute.

b) (less than 1% of clinical trials): A filter dictionary was constructed for side effect keywords that accounted for less than 1% of patients who responded in clinical trials but had no causal relationship with the drug.

c) (additional reported): Filters were pre-built for further reported abnormal reactions after drug marketing.

Network analysis

We used the open-source network analysis visualization software Gephi to visualize the frequency and relevance distribution of keywords (Table 1) [11]. Co-occurrence frequencies are calculated using the document word matrix to express the nodes corresponding to the keywords and the edges indicating the relevance. Frequencies are used as weights connecting nodes and used as edge data. As a visualization method for the nodes, we calculated the value of the power centrality, and the larger the value, the larger the size of the node (size: 3 to 10), and it was set to change from red to blue. Eigenvector centrality is a method of judging important nodes and many connected nodes as important nodes, rather than nodes with many connected nodes being important nodes. Graph Type, which can indicate the direction of edges, is set to Undirected because the directionality between keywords is not considered. We applied the Fruchterman-Reingold algorithm and the Label Adjust option as the visualization method for the edges. The Fruchterman-Reingold algorithm is useful for analyzing relevance between keywords because it appears close between adjacent nodes but separates between non-adjacent nodes. Then apply the Label Adjust option and you will be able to separate the duplicated nodes, creating an easy graph to search for the most important objective: keywords [12].

Extraction clues for drug repositioning

We try to detect a potential candidate symptom for drug repurposing. Search the derived network map to search for keywords that can lead to drug repositioning. Keywords that are determined to be significant in the search process can be used as clues for recreating selective drugs.

RESULTS

Selected the drug

The most important part to consider for drug selection is the ease of data collection, so there should be a lot of data on the Naver Cafe search results of the selected drug. Drugs with a lot of data have a high prevalence rate, which has a common interest, and are characterized by high accessibility without relying on specialized drugs. High blood pressure, a typical adult disease, has doubled from 2021 to 2014, and currently affects 31% of Korean adults. The most commonly prescribed drug in hypertensive patients is angiotensin II receptor antagonist, and the general drug Cozaar-tab containing the drug was selected as the target drug [13].

Collected the dataset

A total of 911 posts were collected through the crawl until August 2022, based on 2008.12.23, which is the permission date of Cozaar-tab. 85% of the posts were collected at the ‘Kidney Disease Patients Meeting’ cafe. As a result of collecting, 778 posts and 4,386 comments were obtained.
The keyword ‘Asthma’, which has a high centrality but is expressed as a disease name, is similar to ‘dyspnea’, which is a side effect of Cozaar-tab. However, the word mismatch prevents filtering by the filter dictionary and can be interpreted as an erroneously derived result. For key-

Data preprocessing
As a result of word tokenization, out of 8,706 deduplicated keywords, we constructed a document word matrix (DTM) with 270 keywords through extract processing. After that, we constructed and applied a filter dictionary of selective drug Cozaar-tab, and reduced 270 keywords to 175 keywords (Figure 2). Through the filter dictionary, 65% of all keywords were able to be used by concentrating on searching for side effects of drugs.

Analyzed results
The visualization results are the same as in Figure 3. It can be seen that the nodes of the keywords stiffening, pain, drowsiness, asthma, excitement, paralysis, and neurosis are relatively large and appear to have a blue color, indicating relatively large centrality. Since keywords with high centrality are often used together with other keywords, they can generally be in the form of modifying keywords. In fact, it can be seen that the keywords with high centrality are not the names of drug-reaction diseases, but rather the symptoms that express the reactions. Symptom keywords that modify disease names can be used more intuitively when studying drug repositioning.
words with the same symptoms and different expression methods, a method of constructing a standardized dictionary and proceeding with additional filtering can be considered.

‘Drowsiness’ among the derived reaction keywords of Cozaar-tab drug can be proposed as a clue for the progress of drug repositioning to sleep-inducing drugs. In fact, diphenhydramine was used as a cold medicine, but it was repositioned as a sleep-inducing agent because of the strong sedative and drowsy side effects of antihistamines [3].

**DISCUSSION**

In this study, we used patterned data on social networks to investigate unknown drug reaction information for drug repositioning. In addition, through network analysis, we placed links between keywords, searched for connections, and sought significance. As a result of analyzing the Naver Cafe data on Cozaar-tab, which is mainly prescribed for hypertensive patients, it was possible to search not only for known drug reactions but also for unknown reaction keywords. This suggests that social network service data may be a good source for obtaining unknown drug reaction data. Therefore, we propose that the drug reaction keywords obtained through the pipeline of this research can be used as clues for drug repositioning research.

This research has the following limitations. First, the scope of research was limited. We limited our analysis to Naver Cafe users, so it’s a bit difficult to generalize to overall general reactions. However, a community where the majority of people can easily and freely share their thoughts and opinions, the study of Naver Cafe data can suggest new research directions for drug repositioning. In addition, the results of this study may be more effective when clinical data such as CDM and EMR are analyzed collectively.

Second, we cannot guarantee that all keywords are reactions to drugs. This is because keywords were extracted without considering causality. However, by embedding whole sentences, checking the relationship between keywords with drugs, and extracting keywords, the quality of keywords could be improved.

Thirdly, there are missing data in the process of keyword extraction. If a specific keyword does not match 100% with the words in the WHO-ART codebook used, it cannot be extracted as a keyword even though they have the same meaning. For example, if the order of tokens is different, such as “weight is loss”, the keyword “weight loss” registered in the codebook cannot be used. Such issues can be addressed by augmenting the data in the WHO-ART codebook to improve the quality of keywords.

**CONCLUSION**

Through this research, we designed a pipeline that establishes the basis for drug repositioning with data from social network services. Also proposed to repositioning Cozaar-tab, a prescription drug for hypertension, as a sleep inducer through the pipeline. In future research, I would like to build the standardized dictionary presented above and advance research to improve the quality of keywords. The derived keywords can then be explored for drug off-label prescription applications through collaboration with clinical experts, and work can be done to further build the filter dictionary. In this way, it is expected to create cases of drug repositioning reflecting the domain knowledge of drug clinical practice.

**ORCID**

Wongyun Lee  https://orcid.org/0000-0002-0426-4527
Seunghee Lee  https://orcid.org/0000-0002-2273-0915
Jong-Yeup Kim  https://orcid.org/0000-0003-1230-9307
Suehyun Lee  https://orcid.org/0000-0003-0651-6481

**REFERENCES**


국문초록
소셜 네트워크 서비스의 데이터를 활용한 약물재창출 단서 추출 파이프라인 연구
이원균¹, 이승희², 김종엽³, 이수현⁴
¹건양대학교병원 헬스케어데이터사이언스센터 연구원, ²건양대학교병원 헬스케어데이터사이언스센터 연구교수, ³건양대학교병원 헬스케어데이터사이언스센터 교수, ⁴가천대학교 IT융합대학 컴퓨터공학과 교수

목적: 본 연구에서는 소셜 네트워크 서비스 이용하여 약물 재창출의 네트워크 분석 기반 파이프라인을 제안하고자 한다.
결과: 코자정의 예상치 못한 키워드 ‘졸음’에서 수면유도제로의 약물 재창출에 대한 근거를 마련하고자 논의하였다.
결론: 이 과정은 향후 임상 오프라벤 검토, 추가 데이터 획득, 네트워크 분석 고도화 등으로 보완하여 데이터 기반의 약물 재창출 연구에 기여할 것으로 기대된다.
주제어: 네트워크 분석, SNS, 약물재창출, 파이프라인