Detection of New Drug Indications from Electronic Medical Records

Tran-Thai Dang[1], Phetnidda Ouankhamchan[1], Tu-Bao Ho[1,2]
[1]Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi City, Ishikawa 923-1292 Japan
[2]John Von Neumann Institute, Vietnam National University at Ho Chi Minh City
Linh Trung, Thu Duc, Ho Chi Minh City, Vietnam
Email: {dangtranthai, s1550203, bao}@jaist.ac.jp

***Abstract*—Drug repositioning – detection of new uses of existing drugs – is an emerging trend in pharmaceutical industry. It essentially is a multiple aspect process of analyzing large-scale heterogeneous data for exploiting advantage of off-targets of the existing drugs. Three kinds of omics, phenomic and drug data are often integrated and used to study drug repositioning. The recent prevalence of electronic medical records (EMRs) makes it become an extremely significant resource of phenomic data for drug repositioning in the post-market stage. However, there is still no generic process and method to this end. This work aims to establish such a process and method. The paper addresses the solution of the first two problems in this complex process.**

I. Introduction

Drug repositioning, also commonly referred to as drug repurposing, has become an increasingly important part of the pharmaceutical industry in recent years [1]. It is defined as the discovery of new possible indications of existing drugs to treat other diseases. For example, aspirin is recently one of the well-known repositioned drugs [2]. Initiating from a research laboratory, aspirin is indicated to treat pain and to reduce fever or inflammation [3]. Lately, aspirin has been discovered to work effectively to prevent cardiovascular disease and colorectal cancer [4].

Developing a new drug through laboratory known as *de novo* R&D approximately costs 359$ millions during a period of 12-years in average [5]. Despite the advances in genomics, life sciences and technology in pharmaceutical industry, the *de novo* drug discovery remained time-consuming and costly, and thus drug repositioning has received much attention as a promising, fast, and cost effective method [6]. As an example, among the 84 drug products introduced to market in 2013, new indications of existing drugs accounted for 20% [7].

In 2011 and 2012, the United Kingdom's Medical Research Council and the US National Center for Advancing Translational Sciences (NCATS), launched large-scale initiatives on drug repositioning, respectively [8]. These pilot programs with participation of major pharmaceutical organizations also promote scientists to conduct creative research on drug repositioning.

However, drug repositioning is an extremely complicated process, a kind of looking for a needle in a haystack. As the drug-disease relationship can be observed in different contexts, drug repositioning can essentially be viewed as a multiple aspect process of mining large-scale heterogeneous data by advanced data analytics methods, aiming to exploit advantage of off-target of the existing drugs. There are notable review articles in the current infancy of drug repositioning [6], [9], [10], [11], [12], [13], [14], [15].

From the literature we can see that the data-driven approach is essential for drug repositioning. On the one hand, the drug repositioning process addresses a very complex relationship between diseases and drugs via the therapeutic targets [16]. That leads to a common framework of multiple databases and integration of the three main resources of (i) genomic data, (ii) phenomic data, and (iii) drug data (i.e., drug chemical compounds). One the other hand, different machine learning methods have always been employed to analyze the above integrated data.

Much work focuses on schemes for integration of multiple databases and interaction among objects represented by those data. In [11], the authors provided a guidance for prioritizing and integrating drug-repositioning methods and tools available in chemoinformatics, bioinformatics, network biology and systems biology. In [17], the authors developed DrugNet that integrates data from complex networks of interconnected drugs, proteins and diseases and applied DrugNet to different types of tests for drug repositioning. In [18], the authors analyzed 'omics' data from genome wide association studies (GWAS), proteomics and metabolomics studies and revealed 992 proteins as potential anti-diabetic targets in human, and 108 of these proteins are verified to be drug targets. In [19], the authors proposed an open source model that supports human-capital development through collaborative data generation, open compound access, open and collaborative screening, preclinical and possibly clinical studies. It is worth noting that the omics data are widely used in pre-market stage of drug development.

There are also a considerable number of papers that focus on exploiting the relation among the data types. A computation method for discovery of new uses of existing drugs is based on the idea that similar drugs are indicated for similar diseases [7]. A new scores produced by large-scale drug-protein target docking on high-performance computing

978-1-5090-4134-3/16/$31.00

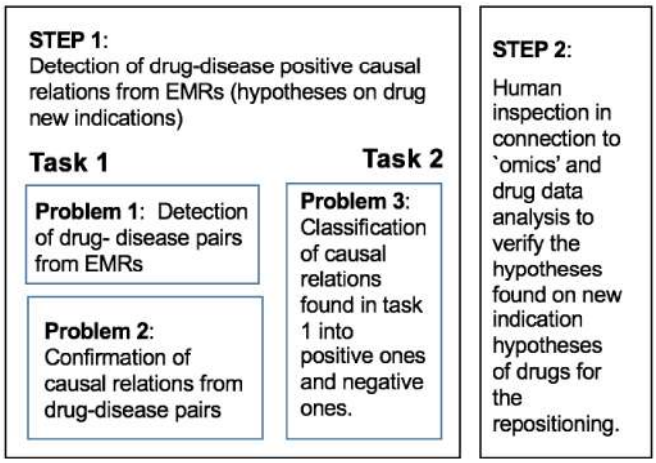

Fig. 1. The process proposed for finding drug new indications from EMRs.

machines [20]. Multiple similarities have been developed to effectively manage multiple integrated databases [21].

Natural language processing (NLP) and text mining are also used in drug repositioning. In [22], the authors used NLP techniques to extract drug indications from structured drug labels. In [23], the authors employed machine learning methods to check off-label drugs from clinical text, Medispan and Drugbank. They detected novel off-label uses from 1,602 unique drugs and 1,472 unique indications, and validated 403 predicted uses. More recent and significant, there are two articles on exploiting electronic medical records (EMRs) for drug repositioning [24], [25]. In [24], the authors used EMRs to study new indications of metformin associated with reduced cancer mortality, and in [25], EMRs are used to repurpose terbutaline sulfate for amyotrophic lateral sclerosis. The clinical text from EMRs in our view will play an extremely important role in drug repositioning, especially in the post-market stage of drug development. However, there is no work so far in the literature addressing a generic process and method on exploiting EMRs for drug repositioning.

Motivated from the lack of such a process and methods for using EMRs in drug repositioning, our work aims to establish a generic process and develop methods for drug repositioning with EMRs. This paper addresses the solution for the first part of the process, i.e., detecting from EMRs the drug-disease pairs that the drug may effect on the disease.

We describe the process and tasks in drug repositioning from EMRs and the proposed method for doing the first task in Section II. Section III describes the experimental evaluation and Section IV concludes the work.

II. Proposed Method

The detection of new indications of drugs from EMRs is a complex process. Our general framework for drug repositioning from EMRs is depicted in Figure 1. It consists of two steps. Step 1 is to detect positive disease-drug causal relations from an EMR as hypotheses of new drug indications, and Step 2 is to verify those hypotheses by human inspection, also by using omics and drug data. Given an EMR, Step 1 consists of two tasks. Task 1 is to detect the causal relations between diseases and drugs in the EMR and Task 2 is to classify those relations into positive and negative ones. The positive causal relations are considered as hypotheses for drug repositioning. We investigate Task 1 by formulating and solving two problems, one is to detect possible pairs of one disease and one drug from that EMR and the other is to determine if there is a causal relation from each of such pairs, it means that if the drug affects on the disease.

This work addresses the Task 1 for drug repositioning from EMRs. Task 2 carrying out by techniques of sentiment analysis in solving Problem 3 that will be investigated in another work.

A. Problems in Task 1

This task is carried out by solving the two following problems:

Problem 1: *Identifying and extracting terms in EMRs that indicate drugs and diseases.*

Problem 2: *Confirming whether there is a relation between an extracted drug and an extracted disease. The relation is known as the drug repositioning or the bad effect of the drug on the disease.*

Essentially, Problem 1 is to recognize the name of drugs and diseases, known as a Name Entity Recognition (NER) problem.

In Problem 2 the relation between drugs and diseases can be described in a bipartite. Denote by U and V two sets of drugs and diseases, respectively, and the chance (strength) of a relation existed between a drug U_i and a disease V_j as the weight w_{ij}. Mostly, each weight w_{ij} is a single value, but if we like to examine the drug-disease associations in multiple perspectives, w_{ij} can be extended into a set $w_{ij} = \{a_1, a_2, ..., a_n\}$ in which each element is a measure according to a perspective. The problem is to appropriately identify w_{ij} that we can base on to precisely confirm the drug-disease associations.

B. Framework of Task 1

In EMR's clinical text, each relation between drugs and diseases is often implicitly mentioned in one or several sentences instead of explicitly mentioning in a formal sentence like in medical articles, and the text in EMRs is almost notes that are written in an informal way. That makes common tools to extract binary relations in a sentence based on syntactic constraints like Reverb [26] become ineffective when applying for EMR's clinical text to detect drug-disease relations. Therefore, to adapt with EMR's clinical text, we develop a statistics-based measure of associations between two entities to determine pairs of drug and disease having a relation. The drug-disease association is measured by considering a large number of patient's clinical notes.

Our proposed framework showed in Figure 2 for detecting drug-disease relations is specified through two phrases: *drug-*

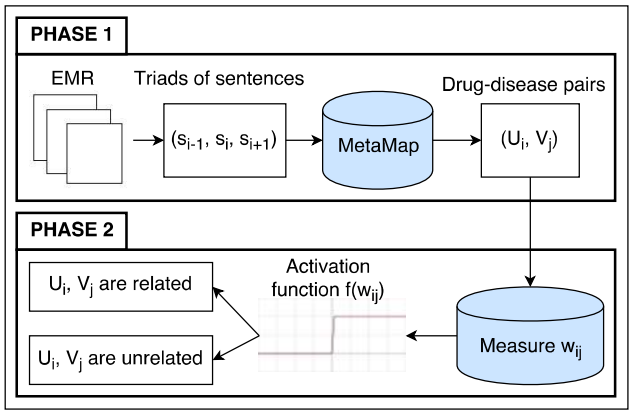

Fig. 2. Our proposed framework to solve problem 1 and 2 in task 1.

disease pairs extraction (phase 1), and *drug-disease relations confirmation* (phase 2).

The purpose of phase 1 is to extract all possible drug-disease pairs (U_i, V_j) mentioned in each discharge summary, doctor daily notes or nurse narratives (note event). Since a drug and its related diseases can appear in different sentences, we need to group these sentences to extract the related drug-disease pairs. To this end, our key assumption is that if a sentence s_i mentions about a drug, the related diseases are often mentioned in s_i or in the neighbor sentences of s_i. Based on this assumption, the drug-disease pairs will be extracted from triads of sentences (s_{i-1}, s_i, s_{i+1}). In addition, the terms indicating drugs and diseases are determined by using MetaMap[1] – a well-know Natural Language Processing (NLP) tool for analyzing biomedical text which gives us the category of each word (semantic type of words).

After extracting the drug-disease pairs in phase 1, in phase 2, for each drug-disease pair we need to confirm whether the corresponding drug and disease are in causal relations or not. This confirmation requires to provide an evidence on possible relations between them. In this case, the evidence is the weigh w_{ij} that characterizes how much U_i and V_j are associated. Estimating an appropriate weight w_{ij} that likely reflects a drug-disease association is a challenge, which is a key point in our work and is presented in detail in subsection II-C. Relying on the estimated weight, we use an activation function $f(w_{ij})$ to classify the drug-disease pairs into two classes "related" and "unrelated". We expected to discover new drug indications in drug-disease pairs belonging to "related" class.

C. Solution for Problem 1 and Problem 2

1) Problem 1: Drug-disease pairs extraction: This phrase consists of extraction of sentence triads and extraction of drug-disease pairs.

In extraction of sentences triads, relying on the assumption mentioned above, a list of drugs under consideration is used to determine sentences s_i that contain the name of those drugs. After that, we consider the previous sentence and the next sentence of s_i to form a triad (s_{i-1}, s_i, s_{i+1}).

The terms indicating drugs and diseases are extracted from the triads of sentences obtained in previous step by using MetaMap [27]. MetaMap is a well-known NLP system that serves to map a given term in a biomedical text to a concept with a corresponding semantic type defined in Unified Medical Language System (UMLS) Metathesaurus. The UMLS incorporates various NLP tools that allow us to break a sentence into phrases and words then map those phrases and words to their semantic types. In our work, after running MetaMap, we select terms with semantic types of "Drug", and "Disease" and form such terms into drug-disease pairs (U_i, V_j).

2) Problem 2: Drug-disease relations confirmation: After extracting pairs (U_i, V_j), we investigate whether U_i and V_j are related or not through estimating the weight w_{ij} that is measured by using Pointwise Mutual Information (PMI) as follows:

$$PMI(U_i, V_j) = \frac{Pr(U_i, V_j)}{Pr(U_i) \times Pr(V_j)} = \frac{c(U_i, V_j) \times N}{c(U_i) \times c(V_j)} \quad (1)$$

where

- $c(U_i, V_j)$, $c(U_i)$, $c(V_j)$ are frequencies of (U_i, V_j), U_i, V_j respectively.
- N is total number of drug-disease pairs extracted from triads of sentences.

$PMI(U_i, V_j) > 0$ if U_i and V_j is associated and vice versa. Therefore, we use a binary step function as an activation function to filter drug-disease pairs to obtain related ones as follows

$$f(w_{ij}) = \begin{cases} 0 & w_{ij} < 0 \\ 1 & w_{ij} \geq 1 \end{cases}$$

Although PMI is an effective statistics-based measure widely used in many problems, in several cases mentioned as below, it shows some drawbacks due to just basing on frequencies $c(U_i, V_j)$, $c(U_i)$ and $c(V_j)$.

- If U_i, V_j are unrelated but co-occur in many times that makes PMI high and leads to lots of redundant drug-disease pairs in the retrieved ones. We consider that as an incorrect suspicion and the precision in this case will be low.
- If U_i and V_j are unrelated, $c(U_i, V_j) \approx c(U_i) \times c(V_j)$ and $c(U_i, V_j), c(U_i), c(V_j) \ll N$, the precision is also low.
- If U_i and V_j are related, but less frequent and $c(U_i, V_j) \ll c(U_i) \times c(V_j)$, the pairs can be left out and the recall will be low.

From the cases of PMI mentioned above, it raises two issues. The first one is how to reduce the unrelated drug-disease pairs in the retrieved ones even though the recall will decrease but we can make the reduction of recall as small as

[1]https://metamap.nlm.nih.gov/

possible. The second one is how to recognize related drug-disease pairs that rarely appear to increase the recall. In the scope this study, we focus on dealing with the first problem.

To remove redundant retrieved drug-disease pairs, we additionally use several constraints to filter the result.

3) Additional constraints for drug-disease relations confirmation: We use constraints of drug-disease frequency or disease-disease relations and PMI together as the weight to eliminate unrelated drug-disease pairs. That means the weight w_{ij} is a set including a measure of the constraint and PMI. Three constraints proposed by us are presented as follows:

- *High Drug-Disease Pair Frequency (constraint 1)*: We will not suspect that the drug and disease are associated if they co-occur less than a predefined threshold η. That means we will eliminate pairs (U_i, V_j) with $c(U_i, V_j) < \eta$.
- *High Disease-Disease Pair Frequency (constraint 2)*: This constraint is based on a concept of comorbidity in medicine. Comorbidity refers to the co-occurrence of several diseases in which some diseases cause the others. We assume that a drug U_i used to treat a disease V_j can affect on another disease V_k which often co-occur with the disease V_j. Before using PMI to discover related drug-disease pairs, we select pairs of related diseases through considering their frequency $c(V_j, V_k)$ that should be greater than a predefined threshold η.
- *Diseases associated with a group of major diseases that a drug is likely related to (constraint 3)*: This constraint is also based on the relations among diseases, but the strategy is different from constrain 2. The idea of this constraint is that a drug is often used to treat some major diseases, and these diseases can cause other diseases. Therefore, the major diseases are known as diseases that have many related ones. We will consider that there is no relation between the drug and diseases which are not associated with the major diseases.
  After using PMI as a criterion for a prior filter, we obtain a preliminary result that drug U_i is suspected to associate with a list of diseases $\mathbf{V} = \{V_j | j = 1, ..., m\}$, and thus we also eliminate unrelated diseases in $\mathbf{V}$. To do so, in the first step, for each V_j in V, we find all related diseases of V_j by considering the co-occurrence frequency of two diseases. In next step, we select k $(k < m)$ diseases with the largest number of their related diseases. We will consider k selected diseases and all their related ones, and eliminate the rest.

III. Experimental evaluation and discussion

A. Experiment design

As mentioned above, the detection of new indications of existing drugs is a complicated process with several steps and involvement of people with different expertise. As this work focuses on the task 1 of the first step in the process, the experiments are designed to evaluate the proposed method performance in their single task and also in the process of detecting novel drug indications from EMRs. The evaluation is carried out according to several perspectives as follows

- Comparison of the proposed method and Reverb in detecting causal relations between drugs and diseases in terms of precision, recall, and F-measure. We run Reverb and our system on the same large dataset extracted from the MIMIC II database [28] then compare their performance by using an annotated test set presented in detail in subsection III-B.
- Investigation on whether three proposed constraints can help to reduce incorrect suspicion of related drug-disease pairs, and examination of how much recall will be reduced.
- Evaluation of the Task 1 solution in the process of new drug indications detection. To do that, we employ the results from pharmaceutical studies related to new indications of drugs conducted by pharmacists, experts, and base on that to confirm how many retrieved drug-disease pairs are probable.

B. The data

The data used for the experiments are "NOTEEVENTS" records of 4000 patients extracted from the MIMIC II database, including discharge summaries, nurse narratives, radiology reports. The records were done pre-processing and separated into sentences.

In the experiment, we investigate 11 drugs often used to treat cardiac diseases and diabetes including Aggrastat, Ativan, Amiodarone, Dilaudid, Vasopressin, Diltiazem, Nitroprusside, Dopamine, Propofol, Lasix, Insulin.

To evaluate the performance of our proposed method and Reverb, we manually created an annotated test set that contains 1172 drug-disease pairs with 3 labels $\{$"0", "1", "2"$\}$. This work was done by basing on available public pharmaceutical literature that contains studies conducted by pharmaceutical experts. The detail of such 3 labels is as follows:

- Label "0" is assigned to unrelated drug-disease pairs, and drug-disease pairs are suspected to have a relation but without any confirmation.
- Label "1" is assigned to related drug-disease pairs which contain original indications of the drug. We base on two well-known resources Drugs.com[2] and DrugBank[3] to determine if these pairs contain the original indication or not. The indications mentioned in these resources are considered original ones.
- Label "2" is assigned to related drug-disease pairs containing new indications of the drug that have already confirmed by at least one study done by pharmaceutical experts. These studies are presented in medical litera-

[2]https://www.drugs.com/
[3]http://www.drugbank.ca/

TABLE I
EXPERIMENTAL RESULTS

Method	P (%)	R (%)	F (%)	R_{new} (%)
Reverb	53.19	5.12	9.35	2.38
PMI without constrains	49.45	73.16	59.01	74.6
PMI + constrain 1 ($\eta = 1$)	54.27	46.93	50.33	45.24
PMI + constrain 2 ($\eta = 1$)	51.05	64.95	57.17	67.85
PMI + constrain 3 ($k = 40$)	52.26	56.97	54.51	59.92

ture that can be obtained in a well-known repository– PubMed[4].

C. Evaluation metrics

The performance of our proposed method and Reverb is evaluated through Precision, Recall, F-measure. We denote numbers of retrieved drug-disease pairs with labels "0", "1", "2" by n_0, n_1, n_2 respectively (the retrieved drug-disease pairs are assigned labels based on the annotated test set). Additionally, numbers of whole drug-disease pairs with labels "1" and "2" in the test set are denoted by N_1 and N_2 respectively. We define the evaluation metrics precision (P), recall (R), F-measure (F) as follows.

$$P = \frac{n_1 + n_2}{n_0 + n_1 + n_2} \tag{2}$$

$$R = \frac{n_1 + n_2}{N_1 + N_2} \tag{3}$$

$$F = 2 \times \frac{P \times R}{P + R} \tag{4}$$

In equation 2, 3, 4, we just investigate related drug-disease pairs that include both pairs with labels "1", "2". Besides, to evaluate our solution for Task 1 in process of detecting new indications of drugs, we also additionally consider the recall of retrieved new indications (R_{new}) as the following.

$$R_{new} = \frac{n_2}{N_2} \tag{5}$$

D. Results

The experimental results when using Reverb and our proposed method in the process of identifying causal relations between drugs and diseases are showed in Table I. For each constraint, we present the result with the most appropriate threshold that gives the best F-measure.

The change of precision, recall when we change the thresholds of the constraints is illustrated in Figure 3. We will base on that to make a comparison among 3 proposed constraints.

E. Discussion

For comparison of the performance between Reverb and our proposed method in the process of identifying causal relations of drugs and diseases, Table I shows that although the precision of Reverb and the proposed method is similar the recall of Reverb is much lower than that of our method. The reason why Reverb gives a very bad recall is that it essentially bases on the part-of-speech patterns containing a main verb which links between two noun/noun phrases to extract binary relations in a sentence, however in EMRs the related drugs and diseases are almost indirectly mentioned in different sentences without linking verbs. Therefore, our proposed method is more appropriate than Reverb in extracting and confirming related drug-disease pairs from EMR data.

As several drawbacks of PMI mentioned above, three constraints are proposed to reduce the incorrect suspicion of related drug-disease pairs. Lines 2-5 of Table I show a improvement when using additionally our proposed constraints to reduce number of unrelated drug-disease pairs blended in the retrieved result. The constraints make precision increase 2-5%.

Although the proposed constraints help to increase of precision, they lead to the significant reduction of recall that is showed in the third column of lines 2-5 of Table I. As the constraints select drug-disease pairs by considering drug-disease or disease-disease pairs which highly frequently co-occur, the related ones but infrequently appear will be left out. It show a drawback of our proposed method that is ineffective in detecting drug indications rarely occurring.

Despite the decrease of recall we expect this reduction is as small as possible. Therefore, we compare 3 proposed constraints to see which one is better to minimize the recall reduction. Figure 3 shows the change of precision and recall when we change the thresholds of each constraint. In constraint 1, when we increase η that means making a tighter restriction of selected drug-disease pairs, the recall rapidly reduces (from 47% to 12%). However, when restricting more tightly in constraints 2 and 3 (increase η in constraint 2 and decrease k in constraint 3), the recall reduce from 64%-42% with constraint 2 and from 60%-42% with constraint 3, and the reduction is much lower than that of constraint 1. Additionally, Table I also shows the higher recall when using constraint 2 and 3. The results show a characteristic of EMR data that in clinical narratives, disease-disease relations are mentioned more frequently than drug-disease relations, so the assumption of basing on disease-disease relations to infer the drug-disease association helps us avoid leaving out related drug-disease pairs that are infrequently mentioned in clinical text. That means constraints 2 and 3 are better than constraint 1 to narrow the recall reduction.

The last column of Table I shows a promising result when using our proposed method to solve Task 1 in process of new drug indications detection. The new drug indications retrieved and confirmed by other studies done by pharmaceutical experts approximately account for from 50%-70% of total number of those annotated in the test set. This result shows a new opportunity for detecting novel drug indications from EMRs by using our proposed method.

IV. CONCLUSION

The paper presents a general framework for drug repositioning based on EMRs in which our initial study concentrates

[4]http://www.ncbi.nlm.nih.gov/pubmed

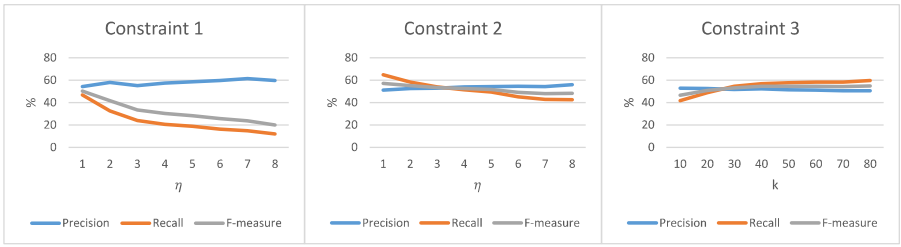

Fig. 3. Investigation of constraints 1,2,3 with different thresholds

on solving two problems of Task 1. We propose a method that essentially bases on PMI-a statistics-based measure to determine drug-disease causal relations with several constraints to improve the precision. This method is more adaptive than syntactic-based methods in detecting drug-disease causal relations on EMRs. The experiments also show that the proposed method is promising to open an opportunity to detect novel drug indications from EMRs. Although this study is still in early stage and requires many improvements in method to achieve higher performance, it forms a groundwork for further studies of EMR-based drug repositioning.

ACKNOWLEDGMENTS

This work is partially funded by Vietnam National University at Ho Chi Minh City under the grant number B2015-42-02.

REFERENCES

[1] M. Barratt and D. Frail, *Drug repositioning: Bringing new life to shelved assets and existing drugs*. John Wiley & Sons, 2012.

[2] K. Banno, M. Iida, M. Yanokura, H. Irie, Y. Masuda, K.and Kobayashi, E. Tominaga, and D. Aoki, "Drug repositioning for gynecologic tumors: a new therapeutic strategy for cancer," *The Scientific World Journal*, vol. 2015, 2015.

[3] Aspirin uses, dosage, side effects & interactions drugs.com. [Online]. Available: "https://www.drugs.com/aspirin.html/"

[4] Cancer.org. (2016) Aspirin and cancer prevention: What the research really shows. [Online]. Available: "http://www.cancer.org/research/acsresearchupdates/cancerprevention/aspirin-and-cancer-prevention-what-the-research-really-shows"

[5] C. B. R. Institute. New drug development process. [Online]. Available: "http://www.ca-biomed.org/pdf/media-kit/fact-sheets/CBRADrugDevelop.pdf"

[6] H. Lee and Y. Kim, "Drug repurposing is a new opportunity for developing drugs against neuropsychiatric disorders," *Schizophrenia research and treatment*, vol. 2016, 2016.

[7] J. Li and Z. Lu, "An integrative approach for discovery of new uses of existing drugs," *Data Science Journal*, vol. 14, 2015.

[8] D. Frail, M. Brady, K. Escott, A. Holt, H. Sanganee, M. Pangalos, C. Watkins, and C. Wegner, "Pioneering government-sponsored drug repositioning collaborations: progress and learning," *Nature Review*, vol. 14, pp. 833–841, 2015.

[9] S. Beachy, S. Johnson, S. Olson, A. Berger *et al.*, *Drug Repurposing and Repositioning: Workshop Summary*. National Academies Press, 2014.

[10] M. Hurle, L. Yang, Q. Xie, D. Rajpal, P. Sanseau, and P. Agarwal, "Computational drug repositioning: From data to therapeutics." *Clinical Pharmacology & Therapeutics*, vol. 93, pp. 335–341, 2013.

[11] G. Jin and S. Wong, "Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines," *Drug discovery today*, vol. 19, no. 5, pp. 637–644, 2014.

[12] G. Wilkinson and K. Pritchard, "In vitro screening for drug repositioning," *Journal of biomolecular screening*, vol. 20, no. 2, pp. 167–179, 2015.

[13] J. Li, S. Zheng, B. Chen, A. Butte, S. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings in bioinformatics*, vol. 17, no. 1, pp. 2–12, 2016.

[14] J. Shim and J. Liu, "Recent advances in drug repositioning for the discovery of new anticancer drugs," *Int J Biol Sci*, vol. 10, no. 7, pp. 654–63, 2014.

[15] T. Ho, L. Le, T. Dang, and S. Taewijit, "Data-driven approach to detect and predict adverse drug reactions," *Current Pharmaceutical Design*, vol. 22, no. 23, pp. 3498–3526, 2016.

[16] J. Dudley, T. Deshpande, and A. Butte, "Exploiting drug–disease relationships for computational drug repositioning," *Briefings in bioinformatics*, 2011.

[17] V. Martinez, C. Navarro, C.and Cano, W. Fajardo, and A. Blanco, "Drugnet: Network-based drug–disease prioritization by integrating heterogeneous data," *Artificial intelligence in medicine*, vol. 63, no. 1, pp. 41–49, 2015.

[18] M. Zhang, H. Luo, Z. Xi, and E. Rogaeva, "Drug repositioning for diabetes based on'omics' data mining," *PloS one*, vol. 10, no. 5, p. e0126082, 2015.

[19] M. Allarakhia, "Open-source approaches for the repurposing of existing or failed candidate drugs: learning from and applying the lessons across diseases," *Drug Des. Dev. Ther*, vol. 7, pp. 753–766, 2013.

[20] M. LaBute, X. Zhang, J. Lenderman, B. Bennion, S. Wong, and F. Lightstone, "Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines," *PloS one*, vol. 9, no. 9, p. e106298, 2014.

[21] P. Zhang, P. Agarwal, and Z. Obradovic, "Computational drug repositioning by ranking and integrating multiple data sources," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 579–594.

[22] K. Fung, C. Jao, and D. Demner-Fushman, "Extracting drug indication information from structured product labels using natural language processing," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 482–488, 2013.

[23] K. Jung, P. LePendu, W. Chen, S. Iyer, B. Readhead, J. Dudley, and N. Shah, "Automated detection of off-label drug use," *PloS one*, vol. 9, no. 2, p. e89324, 2014.

[24] H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, M. Levy, A. Shah, X. Han, X. Ruan *et al.*, "Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 179–191, 2015.

[25] H. Paik, A. Chung, H. Park, R. Park, K. Suk, J. Kim, H. Kim, K. Lee, and A. Butte, "Repurpose terbutaline sulfate for amyotrophic lateral sclerosis using electronic medical records," *Scientific reports*, vol. 5, 2015.

[26] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1535–1545.

[27] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[28] J. Lee, D. J. Scott, M. Villarroel, G. D. Clifford, M. Saeed, and R. G. Mark, "Open-access mimic-ii database for intensive care research," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 8315–8318.