Sequence-based Measure for Identifying Drug-Side Effect Causal Relation from Electronic Medical Records

Tran-Thai Dang[1] and Tu-Bao Ho[1,2]

[1] Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, Japan
[2] John Von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam
{dangtranthai, bao}@jaist.ac.jp

Abstract. The recent prevalence of electronic medical records offers a new way to detect the drug-side effect causality. However, this approach faces with the problem of identifying the likely causal relation between drugs and side effects in a huge space of possible relations in which many relations are not causal ones, but frequently observed. In existing work, the likely causal relation is almost detected by using frequency-based measures of drug-side effect pair co-occurrence, but the accuracy is rather low due to the frequent co-occurrence of non-causal pairs. Our key assumption on the causality of the drugs and side effects is that the causality occurring at a medication event has an association with the therapy history of this event. The assumption is employed as a constraint in the proposed sequential-based model named Medication Therapy Progress-based Model (MTPM). Experiments show a significant improvement of accuracy from 4% to 9% when comparing MTMP and existing methods, as well as reflect the likelihood of the assumption.

1 Introduction

Drug side effect (also called adverse drug reactions) can be understood as undesirable effect, reasonably associated with the use of the drug that may occur as a part of the pharmacological action of a drug or may be unpredictable in its occurrence. Drug side effect detection plays an essential role in drug safety. The side effect can be caused by various reasons such as over dosing, the interaction between the drug and off-target, or the interaction between drugs. Before being approved for using, a drug has to go through a series of clinical trials to evaluate expected indications and its possible side effects. Such trials are often conducted under ideal and controlled circumstances, called explanatory clinical trials, that can only test efficacy of the drug but not its effectiveness.

The effectiveness of a drug is evaluated by pragmatic clinical trials [24], i.e., to see how well the drug works in the "real world". However, pragmatic clinical trials (PCT) are much more difficult to do [10]. Carrying out pragmatic clinical trials to detect drug side effects is basically analyzing textual data coming from patient spontaneous reports, reports in social network, and electronic medical record (EMRs) [12], [24], [30]. The clinical texts from EMRs contain almost all the facts about drug effects observed under the real condition of a huge patient cohort. EMRs are well recognized as a precious resource for pragmatic clinical trials. The difficult problem is to recognize the side effect and to assess causality between drugs and side effects [9].

EMR data has considerable advantages in assessing the drug-side effect causality in comparision with the patient spontaneous reports and social media data. In the force of pragmatism in clinical research and from the EMR opportunity, there have been research, even still in its infancy, to use electronic medical records for pragmatic clinical trials, under the abbreviation EMRPCT [8], [25], [29]. In [3], the authors present and analyze the theoretical advantages and disadvantages, the ethical and regulatory aspects of EMRPCT, as well as prospects of EMRPCT in drug effectiveness study. Typically, EMR data has two main properties, one is longitude and the other is heterogeneity [19] that bring a new opportunity in clinical research as well as pose many challenges in analyzing and mining EMR clinical text [13]. EMR clinical text is less bias than patient reports or social networks because it is captured more objectively and sufficiently by medical experts. Moreover, EMR data includes more diverse populations and rare diseases.

Several work has been pursued to detect drug-side effect causality from EMRs. They commonly follow a two-step framework, the first one is to recognize two sets of named entities for drugs and effects, and the second one is to detect drug-side effect causal pairs from those two sets. Note that the set of side effects is subset of the effect set extracted from the clinical text. The common point of those work is to investigate the causality of drug-side effect pairs based on observing their co-occurrence in documents, quantified by frequency-based measures. In [20], Liu *et al.* measured the association between side effects and statin drugs by using log-likelihood ratio based on the proportion between the number of statin drug reviews and non-statin drug reviews. In [5], [21], [28], χ^2 statistics was commonly used to confirm the association between drugs and effects. In [27], Wang *et al.* used Pairwise Mutual Information (PMI) for this target. Roitmann *et al.* [23] considered the morbidity caused by drugs through investigating the co-occurrence of adverse events. Besides, the drug-effect causality is also represented in the form of association rules, and strength of rules is supported by well-known measures such as RR [11], support, confidence, leverage [2], [15], [16], [31]. In addition, in [26], a hypothesis of an association between side effects and therapeutic indication was investigated by using a predictive model.

The methods in the above mentioned work were conducted on documents from patient report systems, social networks, and EMRs, but while being appropriate for the first two kinds of textual data they do not work well on EMR clinical data. The main reason lies in the difference between those data types. In patient reports, the causal relation is more explicit, even is mentioned directly by patients. As such relation comes from the feeling or perception of the patients, so the certainty in identifying the causality is fairly high. On the other hand, the clinical notes in EMRs are basically narratives, written in an objective way with a high proportion of noun for describing observations [7], so the causality almost has not mentioned in such text causing uncertainty in the detection. In addition, there is a huge space of possible relations due to multiple use of drugs and objectively noting observations, in which many non-causal drug-side effect pairs coincidentally observed in a high frequency that makes the pairwise association measures need to be adapted for EMR clinical text.

2 Problem Formulation

Side effects can be caused by a single drug, or interaction among multiple drugs. Identifying side effects caused by multiple drugs is more complicated than that caused by single drug, but is promising to discover new effects beyond human knowing. This problem can be simplified by converting to the problem of detecting side effects caused by the single drug because the side effects of a drug combination can be considered side effects caused by each drug in this combination.

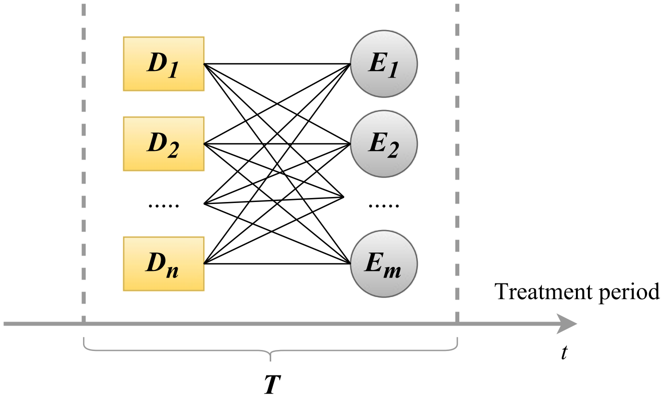

Fig. 1. Bipartite graph represents all possible associations between drugs and side effects observed in a time window T, in which, function $w(D_i, E_j)$ is weight of an association link that measures the association strength between the drug D_i and the effect E_j.

In EHRs/EMRs, all possible temporal associations between drugs and observed side effects are often investigated within a identified time window T during a therapy period to find likely ones [15], [16]. These associations can be represented by a bipartite graph illustrated in Figure 1 with the function $w(D_i, E_j)$ used to measure the association strength between the drug D_i and the effect E_j. The problem of identifying/selecting likely causal drug-side effect pairs can be viewed as a problem of ranking all candidates of pairs in the bipartite graph according to their association strength. Therefore, the groundwork for effectively ranking is to find a measure of drug-side effect relation strength that well reflects the real causal relation, which is the objective of our work.

3 Existing Methods of Measuring Drug-Side Effect Causal Relation Strength

Most of researches on identifying drug-side effect causal relations so far present the causal relations between drugs and side effects in form of temporal association rules $D_i \xrightarrow{T} E_j$ with various pairwise statistical association measures for quantifying strength of rules [15], [16], [22].

The association strength can be estimated through several kinds of measures such as Confidence ($conf$), Leverage (lev) [15], [16], χ^2 test [5], and Relative Reporting Ratio (RR, which basically is similar to Pointwise Mutual Information) [11] as follows:

$$conf(A \xrightarrow{T} C) = \frac{supp(A \xrightarrow{T} C)}{supp(A \xrightarrow{T})} \tag{1}$$

where $supp(A \xrightarrow{T})$ is proportion of $T-$constrained sub-sequences containing A.

$$lev = supp(A \xrightarrow{T} C) - supp(A \xrightarrow{T}) \times supp(\xrightarrow{T} C) \tag{2}$$

$$RR = N \times S(A \cup B)/S(A)S(B) \tag{3}$$

where N is total number of records in the data, $S(A \cup B)$, $S(A)$, $S(B)$ are support measure of $A \cup B$, A, B, respectively.

4 Sequence-based Measuring Drug-Side Effect Causal Relation Strength

In the general framework of detecting drug-side effect causal relation, before estimating the temporal causal association strength between drugs and side effects to select likely causal pairs, a hospitalization period needs to be divided into time windows T, then two sets of names entities for drugs and effects within such windows need to be identified that is presented in Subsection 4.1. In the scope of our study, we concentrate on measuring the causal relation between drugs and their side effects that is intensively mentioned in Subsections 4.2, 4.3, 4.4.

4.1 Forming Time Window in Hospitalization Period from Electronic Medical Records

Our work was conducted on a practical electronic medical record databases named MIMIC-III (Medical Information Mart for Intensive Care III)[1] [17]. This

[1] https://mimic.physionet.org

database contains prescriptions, and clinical notes of a large number of patients during their hospitalization, attached with timestamps when the drugs are prescribed and the clinical notes are recorded.

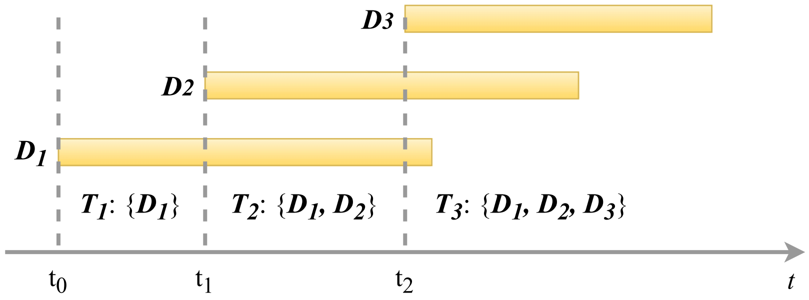

Fig. 2. An example of determining time windows based on information of starting and ending time of drug usage given in EMRs. The orange bars indicate the drug usage period, t_0, t_1, t_2 are starting dates corresponding to the drugs D_1, D_2, D_3, respectively.

Splitting a patient's hospitalization period into time windows is based on starting time of drugs used during the treatment period that is illustrated in Figure 2. A drug is determined to belong to a window if the time interval bounded by this window is in the drug usage period. For example, in Figure 2, in the window T_1 only the drug D_1 is prescribed, then in T_2, the drug D_2 is started to use with D_1, and the last window, all three drugs are prescribed together. After forming time windows, the clinical notes are also mapped to their corresponding window by the time of the notes creation.

After mapping the clinical notes, the set of drug effects is determined by extracting words, phrases expressing symptoms, abnormalities from these clinical notes using MetaMap[2][1]. MetaMap is a well-known Natural Language Processing system for analyzing biomedical text based on Unified Medical Language System (UMLS) Metathesaurus. Two main functions of MetaMap are medical terminology recognition and category (often called semantic type) identification. In order to identify effects in clinical text, we use four semantic types including "Acquired Abnormality" and "Finding" and "Sign or Symptom".

4.2 Assumption about Sequential Association among Drug Effects

In pharmaceutical science, drug is essentially a chemical compound, and drug target is considered as a mass of protein molecules including receptors that receive chemical signal from outside a cell. Due to being protein, the drug target is associated with observed diseases, symptoms which are called phenotype in general [14]. To understand about the mechanism of drug effects, we briefly introduce some relevant biological concepts.

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

- Transcription factor: A protein required to bind to regulatory region of DNA (Deoxyribonucleic acid), and helps to translate "genetic message" in DNA into RNA (Ribonucleic acid) and protein.
- Regulatory region of DNA: A region in the DNA sequence that needs specific proteins to turn it on or sometimes off.
- Gene expression: A process by which information from a gene is used to synthesize protein.

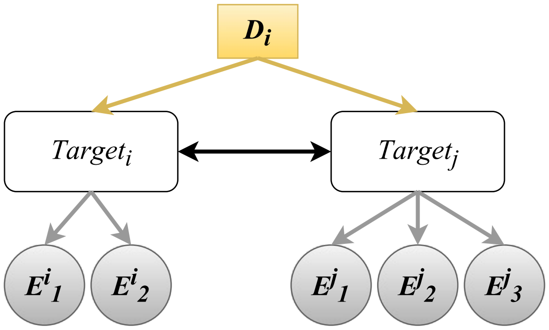

Fig. 3. Assumption about the association among side effects due to the relation with the same target, and the target interaction.

When drugs come in the body, they activate transcription factors, then the transcription factors can bind to regulatory regions of DNA. The DNA changes its status that leads to the gene expression process taking place to change RNA and protein. The change of protein causes phenotype exposure known as diseases, or symptoms.

Besides the interaction between drugs and drug targets, the drug targets also interact with each other due to the protein-protein interaction. Commonly, protein-protein interaction can be understood as physical contacts between proteins that occur in cell or in living organism [6]. The physical contacts mean the functional sharing between proteins. Therefore, the change in functions of a protein can lead to the change in functions of the others.

Relying on the interaction between drugs and targets, and the interaction between the targets which essentially is protein-protein interaction, we make an assumption that there exists the potential association between side effects because of two following reasons:

1. The exposing side effects may has a relation with the same target. This is illustrated in Figure 4.2 that all elements in the set of $\{E^i_1, E^i_2\}$ or the set of $\{E^j_1, E^j_2, E^j_3\}$ may has an association with each other because of relating to the same $Target_i$, $Target_j$, respectively. Intuitively, this reason explains the co-occurrence of side effects in a same family. For example, both respiratory

tract infection, and rhinitis, which are diseases of respiratory system, are side effects of Salbutamol (the drug used to treat asthma)[3].

2. The side effects associated with different targets may have the relation with each other due to the target-target interaction. In Figure 4.2, as $Target_i$, $Target_j$ are interactive, so there may exist an association between elements in their two corresponding effect sets. For practical example, we observe that headache, and fever lie in the list of Salbutamol's side effects. Excluding the reason that the drug impacts directly on the brain, another reason can be considered that respiratory side effects cause breathing difficulty that leads to the headache and fever.

Since the duration of drug action depends on several factors such as the amount of drug given (doses), the pharmaceutical preparation, the reversibility of drug action, the half-life of the drug, the slope of the concentration-response curve, the activity of metabolites, the influence of disease on drrug elimination [4], and different time of taking drug, the co-occurrence of associated side effects is not simultaneous. That means the observation of associated side effects is sequential.

Assumption: *There exists the potential sequential association among side effects of a drug.*

4.3 Inspiration of Sequence-based Drug-Side Effect Causal Relation Suspicion

The medication treatment in the Intensive Care Unit (ICU) often gradually become more complicated due to the appearance of additional diseases (comorbidity) requiring more drugs used. The increase of number of drugs used pulls the increase of number of side effects, which make the causal relation identification become more difficult due to the huge number of possible candidates. This makes most of previous work become ineffective in detecting causal relation, particularly in detecting the low frequency or rare relations [22]. The assumption about the association between side effects inspires us to supplementally exploit the relation between side effects observed in historical medication events with the current ones for reducing the uncertainty in identifying. The idea of sequence-based drug-side effect causal relation suspicion is illustrated in Figure 4.

Figure 4 shows that assuming the side effect observed in the time window T_1 is properly caused by the drug D_i, if the effect in the window T_2 strongly related to the previous proper one, it will be suspected to be the side effects caused by the drug D_i.

4.4 Model

In this subsection, we introduce our proposed sequence-based measure to quantify the strength of drug-side effect causal relation that bases on the assumption

[3] http://sideeffects.embl.de/drugs/2083/

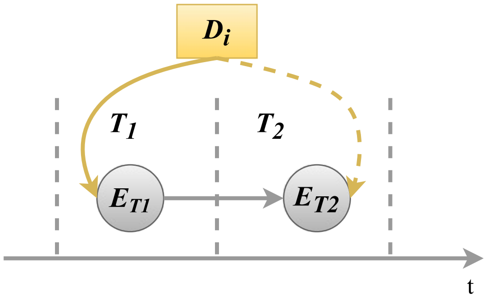

Fig. 4. Idea of Sequence-based drug-side effect causal relation suspicion. The orange and gray solid lines indicate the relations between the drug and the previous effect and between the current effect and the previous one, respectively, which are assumed to be proper. The orange dashed line indicate the relation between the drug and the current effect that is suspected to exist.

about the sequential association between side effects. Identifying causal drug-side effect pairs includes two steps:

1. Identifying likely causal drug-side effect pairs in a hospitalization.
2. Aggregating likely causal drug-side effect pairs observed in several hospitalizations.

Identifying likely causal drug-side effect pairs in a hospitalization We consider the prescription and clinical notes of a patient hospitalization h, in which D is the set of all used drugs and E is the set of all side effects. The hospitalization period is divided into time windows, mentioned in Subsection 4.1. For each drug D_i, n_i is the number of time windows belonging to this drug usage period. We form all possible candidate for selecting likely drug-side effect pairs in n_i windows of the drug D_i, then define the measure of causal relation strength for each candidate $w_h(D_i, E_{jk})$ where $1 \leq i \leq |D|$, $1 \leq j \leq |E|$, and $1 \leq k \leq n_i$ by a recursive function as below:

$$w_h(D_i, E_{jk}) = \begin{cases} log\Big(P(D_i|E_{jk})\Big) & \text{if } k = 1 \\ Q \quad \text{if } 2 \leq k \leq n_i \end{cases} \tag{4}$$

where Q is defined as following:

$$Q = \frac{1}{k} \times \left(log\Big(P(D|E_{jk})\Big) + log\Big(P(E_{jk}|E_{j(k-1)})\Big) + w_h(D_i, E_{j(k-1)}) \right)$$

The probability $P(D_i|E_{jk})$ called emission probability measures the association between the side effect and the drug within the window k, and $P(E_{jk}|E_{j(k-1)})$ is called transition probability that measure the association between side effects observed in two consecutive windows. The emission and transition probability are estimated, respectively, as follows:

$$P(D_i|E_{jk}) = \frac{count(D_i, E_{jk}) + \lambda}{count(E_{jk}) + \lambda \times |E|}$$

$$P(E_{jk}|E_{j(k-1)}) = \frac{count(E_{j(k-1)}, E_{jk}) + \lambda}{count(E_{j(k-1)}) + \lambda \times |E|}$$

where $count(D_i, E_{jk})$, $count(E_{jk}, E_{j(k-1)})$, $count(E_{j(k-1)})$ are number of patients taking the drug D_i and the effect E_{jk} observed, number of patients that both E_{jk} and $E_{j(k-1)}$, and only $E_{j(k-1)}$ are observed, respectively. The predefined constant λ is Laplacian smoothing coefficient that often takes the value of 0.1.

We would like to make the cumulative process in estimating $w_h(D_i, E_{jk})$ smooth to avoid the problem of imprecisely estimating transition and emission probability because of the coincident observation of side effects caused by multiple drugs. The smooth means no rapid change in value of the function w_h between two consecutive windows. For this purpose, we add the smoother $\frac{1}{k}$.

Clearly, the recursive function shows that if the effect in the previous window is causally related to the considering drug, and the effect in the current window is strongly associated to it, the current effect is more suspected to has relation to the drug. That reflects our idea mentioned in Subsection 4.3

As the assumption about the sequential association among side effects of a drug, so for each drug D_i, we find an effect sequence $(\hat{E}_1, \hat{E}_2, ..., \hat{E}_{n_i})$ that maximizes the value of $w(D_i, E_{jn_i})$, which is illustrated in Figure 5. The value of $w(D_i, E_{jn_i})$ is the cumulative value of the sequence that measures how strongly the effects in this sequence are related to the drug. Therefore, the side effects in the selected sequence are identified to have causal relation with the drug in the hospitalization h. The most likely sequence is discovered by using Viterbi algorithm.

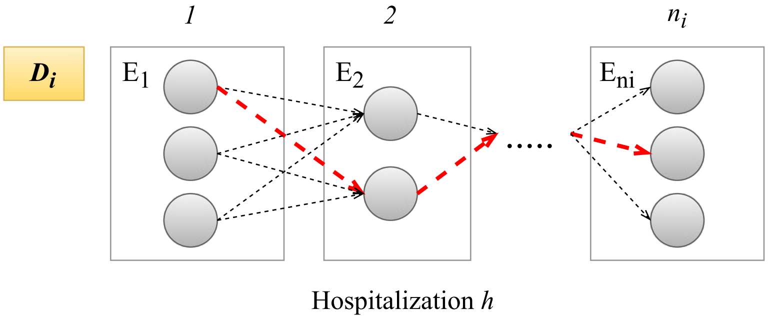

Fig. 5. Detecting the sequence of side effects in n_i windows that has the strongest relation to the drug D_i (lies on red dashed line).

Algorithm 1: Viterbi Algorithm

```
$best\_score = \{\}$
$back\_trace = \{\}$
for $k := 1$ $to$ $n_i$ do
    if $k == 1$ then
        for $x := 1$ $in$ $|E_k|$ do
            Compute $w(D_i, E_{xk})$ according to Eq.4
            $best\_score[D_i, E_{xk}] = w_h(D_i, E_{xk})$
            $back\_trace[D_i, E_{xk}] = None$
    else
        for $x := 1$ $in$ $|E_k|$ do
            for $y := 1$ $in$ $|E_{k-1}|$ do
                Compute all values of $w_h(D_i, E_{xk})$ according to Eq. 4 with
                  different value of $w_h(D_i, E_{y(k-1)})$ then store the results in
                  an array $W$
                $max\_val = max(W)$
                $ymax = W$.index($max\_val$)
            $best\_score[D_i, E_{xk}] = max\_val$
            $back\_trace[D_i, E_{xk}] = E_{ymax(k-1)}$
for $x := 1$ $in$ $|E_{n_i}|$ do
    Select $\hat{E}_{n_i}$ having maximum value of $w_h(D_i, E_{xn_i})$
Get the sequence $(\hat{E}_1, \hat{E}_2, ..., \hat{E}_{n_i})$ using $back\_trace$
```

Aggregating likely causal drug-side effect pairs observed in several hospitalizations The drug-side effect pair (D_i, E_j) can be observed in several hospitalizations H of different patients with different values of $w_h(D_i, E_j)$ where $h \in H$, so to make the final decision whether there exists the causal relation between the drug and the side effect, we aggregate all values of $w_h(D_i, E_j)$ by taking their maximum value.

$$w(D_i, E_j) = \max_{h \in H} \Bigg(w_h(D_i, E_j) \Bigg)$$

5 Experimental Evaluation

5.1 Experimental Design

The data set used for the experiments is MIMIC III (Medical Information Mart for Intensive Care III) briefly mentioned in Subsection 4.1. This data set is large and freely accessible that contains over 40,000 patients who stayed in the Beth Israel Deaconess Medical Center between 2001 and 2012 [17]. It includes various information of demographics, laboratory test, medication events, clinical notes.

For the scope of this study, we used prescriptions and clinical notes for the experiments.

From the MIMIC III database, we exported the prescriptions and clinical notes of 10,000 patients, in which, we detect causal drug-side effect pairs in randomly selected 50 patients by collating with the rest which is used to estimate transition, emission probabilities in our proposed model. The exported raw data was pre-processed by the mechanism mentioned in Subsection 4.1. We select 49 drugs to detect their side effects using proposed model. The performance of the model is evaluated through checking how many causal drug-side effect pairs that are confirmed by SIDER[4] [18] in the retrieval pairs.

We compare the performance of our proposed model with existing methods mentioned in Section 3. The key point is to investigate the quality of association measures used in those methods in reflecting the real drug-side effect causal relation. In previous work, estimating the value of probabilities such as *supp*, *conf* was carried out in a different way from our method, so for fairly comparison, we make a consensus of probability computing which is based on the proportion between the number of patients presenting the relation or property over the total patients. That means we count number of patients whom the drug-effect pairs, effect-effect pairs are observed on, instead of counting the frequency of these pairs mentioned in the clinical text.

5.2 Evaluation Metrics

In this study, we evaluate the performance of the methods in identifying drug-side effect causal relation by Precision K ($Prec_K$) which is defined as the fraction of known side effects occurring in the top K ones of the list returned by each method for a specific drug [22].

$$Prec_K = \frac{\sum_{i=1}^{K} y(i)}{K}$$

where $y(i) = 1$ if the i^{th} side effect is the proper one, and is 0 for otherwise.

5.3 Experimental Results and Discussion

Identifying drug-side effect causal relation in electronic health records or electronic medical records is a challenging problem. The solution for this problem so far is still in early stage that just used conventional statistical measures to directly estimate the strength of drug-side effect relation, which mostly produces low performance. For example, in [16], the authors used leverage measures to detect causal drug-effect pairs in the Queensland Linked dataset, and got the $Prec_{10}$ is about 0.313.

In order to investigate the likelihood of the proposed assumption about the sequential association between side effects appearing in a hospitalization as well

[4] http://sideeffects.embl.de

as the effectiveness of the sequence-based measure utilization for solving this problem, we make a comparison between the proposed method and existing methods with multiple values of K that is showed in Table 1.

Table 1. Performance comparison between sequence-based method and existing methods in identifying drug-side effect causal relation

Method	$Prec_5$	$Prec_{10}$	$Prec_{15}$	$Prec_{20}$	$Prec_{25}$	$Prec_{30}$
RR	0.331	0.33	0.33	0.337	0.333	0.339
$conf$	0.403	0.375	0.386	0.387	0.389	0.39
lev	0.373	0.337	0.343	0.343	0.339	0.335
χ^2 test	0.373	0.346	0.356	0.367	0.369	0.363
Sequence-based measure	**0.437**	**0.447**	**0.439**	**0.439**	**0.433**	**0.427**

Equation 4 shows that the function $w_h(D_i, E_{jk})$ with $k = 1$ (without previous windows)

6 Conclusion

This paper presents our proposed model called Medication Therapy Progress-based Model to recognize drug-side effect causality that is based on the assumption about the association between a drug-side effect causality occurring in a medication event and the therapy history of this event. The experiment shows that the proposed model as well as the assumption can improve the recognizing accuracy and provide an effective score for highlighting causal pairs and distinguishing the causal pairs from non-causal pairs. However, the model has a drawback that the probabilities in the model are estimated from a sample of patients without integrating medical knowledge that is not good for giving the precise probabilities and makes the improvement is not significant. Thus, in the future work, we target to exploit more constraints and incorporate domain knowledge make a significant improvement.

References

1. Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
2. Adrian Benton, Lyle Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E Leonard, and John H Holmes. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6):989–996, 2011.
3. A.J. Carcasa, F.A. Santos, L.S. Perrucac, and R. Dal-Ree. Electronic medical record in clinical trials of effectiveness of drugs integrated in clinical practice. *Medicina Clinics*, 145(10):452–457, 2015.

4. SG Carruthers. Duration of drug action. *American family physician*, 21(2):119–126, 1980.

5. Elizabeth S Chen, George Hripcsak, Hua Xu, Marianthi Markatou, and Carol Friedman. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. *Journal of the American Medical Informatics Association*, 15(1):87–98, 2008.

6. Javier De Las Rivas and Celia Fontanillo. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, 2010.

7. Yihan Deng, Matthaeus Stoehr, and Kerstin Denecke. Retrieving attitudes: Sentiment analysis from clinical narratives. pages 12–15, 2014.

8. H.F. Elkhenini, K.J. Davis, N.D. Stein, J.P. New, M.R. Delderfield, M. Gibson, J. Vestbo, A. Woodcock, and N.D. Bakerly. Using an electronic medical record (emr) to conduct clinical trials: Salford lung study feasibility. *BMC Medical Informatics and Decision Making*, 15, 2015.

9. Andreea Farcas and Marius Bojita. Adverse drug reactions in clinical practice: a causality assessment of a case of drug-induced pancreatitis. *J Gastrointestin Liver Dis*, 18(3):353–8, 2009.

10. I. Ford and J. Norrie. Pragmatic trials. *The New England Journal of Medicine*, 375(5):454–463, 2016.

11. Rave Harpaz, Krystl Haerian, Herbert S Chase, and Carol Friedman. Statistical mining of potential drug interaction adverse effects in fdas spontaneous reporting system. *AMIA Annu Symp Proc*, pages 281–285, 2010.

12. T.B. Ho, L. Le, T.T. Dang, and T. Siriwon. Data-driven approach to detect and predict adverse drug reactions. *Current Pharmaceutical Design Journal*, 22(123):3498–3526, 2016.

13. George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

14. Peter Imming, Christian Sinning, and Achim Meyer. Drugs, their targets and the nature and number of drug targets. *Nature reviews Drug discovery*, 5(10):821–834, 2006.

15. Yanqing Ji, Hao Ying, Peter Dews, John Tran, Ayman Mansour, Richard E Miller, and R Michael Massanari. An exclusive causal-leverage measure for detecting adverse drug reactions from electronic medical records. *Fuzzy Information Processing Society (NAFIPS), 2011 Annual Meeting of the North American*, pages 1–6, 2011.

16. Huidong Jin, Jie Chen, Hongxing He, Chris Kelman, Damien McAullay, and Christine M O'Keefe. Signaling potential adverse drug reactions from administrative health databases. *IEEE Transactions on knowledge and data engineering*, 22:839–853, 2010.

17. Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

18. Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

19. Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. pages 705–714, 2015.

20. Jingjing Liu, Alice Li, and Stephanie Seneff. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. *Proceedings of First International Conference on Advances in Information Mining and Management (IMMM), Barcelona, Spain*, pages 23–29, 2011.
21. Mei Liu, Eugenia Renne McPeek Hinz, Michael Edwin Matheny, Joshua C Denny, Jonathan Scott Schildcrout, Randolph A Miller, and Hua Xu. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *Journal of the American Medical Informatics Association*, 20(3):420–426, 2013.
22. Jenna Reps, Jonathan M Garibaldi, Uwe Aickelin, Daniele Soria, Jack E Gibson, and Richard B Hubbard. Comparing data-mining algorithms developed for longitudinal observational databases. pages 1–8, 2012.
23. Eva Roitmann, Robert Eriksson, and Søren Brunak. Patient stratification and identification of adverse event correlations in the space of 1190 drug related adverse events. *Frontiers in physiology*, 5, 2014.
24. Tjeerd-Pieter van Staa, Ben Goldacre, Martin Gulliford, Jackie Cassell, Munir Pirmohamed, Adel Taweel, Brendan Delaney, and Liam Smeeth. Pragmatic randomised trials using routine electronic health records: putting them to the test. *Bmj*, 344:e55, 2012.
25. T.P. van Staa, L. Dyson, G. McCann, S. Padmanabhan, R. Belatri, B. Goldacre, J. Cassell, M. Pirmohamed, D. Torgerson, S. Ronaldson, J. Adamson, A. Taweel, B. Delaney, S. Mahmood, S. Baracaia, T. Round, R.Fox, T. Hunter, M. Gulliford, and L. Smeeth. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technology Assessment*, 18(43):1–141, 2014.
26. Fei Wang, Ping Zhang, Nan Cao, Jianying Hu, and Robert Sorrentino. Exploring the associations between drug side-effects and therapeutic indications. *Journal of biomedical informatics*, 51:15–23, 2014.
27. Xiaoyan Wang, George Hripcsak, and Carol Friedman. Characterizing environmental and phenotypic associations using information theory and electronic health records. *BMC bioinformatics*, 10(9):S13, 2009.
28. Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337, 2009.
29. K. Yamamoto, E. Sumi, T. Yamazaki, K. Asai, M. Yamori, S. Teramukai, K. Bessho, M. Yokode, and M. Fukushima. A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. 2:1–10, 2012.
30. Keiichi Yamamoto, Eriko Sumi, Toru Yamazaki, Keita Asai, Masashi Yamori, Satoshi Teramukai, Kazuhisa Bessho, Masayuki Yokode, and Masanori Fukushima. A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. *BMJ open*, 2(6):e001622, 2012.
31. Christopher C Yang, Ling Jiang, Haodong Yang, and Xuning Tang. Detecting signals of adverse drug reactions from health consumer contributed content in social media. *Proceedings of ACM SIGKDD Workshop on Health Informatics*, 2012.