Mechanism-based Pharmacovigilance over the Life Sciences Linked Open Data Cloud

Maulik R. Kamdar, M.Tech., Mark A. Musen, M.D., Ph.D.
Center for Biomedical Informatics Research, Stanford University, CA, USA

Abstract

Adverse drug reactions (ADR) result in significant morbidity and mortality in patients, and a substantial proportion of these ADRs are caused by drug–drug interactions (DDIs). Pharmacovigilance methods are used to detect unanticipated DDIs and ADRs by mining Spontaneous Reporting Systems, such as the US FDA Adverse Event Reporting System (FAERS). However, these methods do not provide mechanistic explanations for the discovered drug–ADR associations in a systematic manner. In this paper, we present a systems pharmacology-based approach to perform mechanism-based pharmacovigilance. We integrate data and knowledge from four different sources using Semantic Web Technologies and Linked Data principles to generate a systems network. We present a network-based Apriori algorithm for association mining in FAERS reports. We evaluate our method against existing pharmacovigilance methods for three different validation sets. Our method has AUROC statistics of 0.7–0.8, similar to current methods, and event-specific thresholds generate AUROC statistics greater than 0.75 for certain ADRs. Finally, we discuss the benefits of using Semantic Web technologies to attain the objectives for mechanism-based pharmacovigilance.

1 Introduction

Pharmacovigilance methods are used to detect unanticipated adverse drug reactions (ADR) that manifest due to the intake of drugs by patients. These ADRs are often not detected during the clinical trials of the corresponding drugs. A majority of these ADRs are caused by polypharmacy, a situation where multiple concomitant drugs are administered to one patient in a short span of time to treat multiple medical conditions.[1] These drugs may interact with each other through several different underlying biological mechanisms.[2] Drug–drug interactions (DDI) due to polypharmacy are potentially avoidable, if detected early.[3] ADRs are the 4^{th} leading cause of death ahead of diabetes, AIDS, and pneumonia.[4] ADRs often result in the hospitalization or serious injury of more than 2 million individuals in the United States, with more than 100,000 deaths annually.[5] The costs of drug-related morbidity and mortality in the United States alone were estimated to be US$177.4 billion in 2000, and have been rising ever since.[6]

Pharmacovigilance methods often use data from Spontaneous Reporting Systems, such as the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS),[7] or electronic medical records.[8] These methods have inferred new DDIs and the ADRs that manifest on the account of those interactions (e.g., $Vioxx \rightarrow Heart\ Attack$ and $Aspirin + Warfarin \rightarrow Bleeding$). However, these studies do not systematically demonstrate how the drugs interact within the biological system of the patient, leading to a particular adverse reaction. *Mechanism-based pharmacovigilance* can lead to the inference of newer DDIs and ADRs, and can also provide a better understanding of the underlying biological mechanisms behind the DDIs.[9] Moreover, this understanding can lead clinicians to prescribe drugs that can treat the same medical conditions in a patient while minimizing the risk of DDIs due to different mechanisms of those drugs. The objectives of mechanism-based pharmacovigilance can be attained through the development of network-based approaches of integrative pharmacology, often termed *systems pharmacology*.[9] These approaches rely on an exhaustive *systems network*, that possesses knowledge of the drug-induced perturbations of the physiological functions in a biological system as well as knowledge of the underlying biological interactions.

However, the data and knowledge to generate such a network exists in several databases and knowledge bases that may be fragmented across the Web. These sources, if available for download, may: *i)* use varying schemas to structure the data, *ii)* use different entity notations (e.g., proteins referenced using HGNC[10] or KEGG[11] identifiers), and *iii)* use different formats for storage (e.g., XML, CSV, etc.). An *ad hoc* integration approach involving downloading and integrating each source independently, and reconciling similar entities, is non-trivial, non-scalable and is often redundant for different tasks. Hence, the objectives of mechanism-based pharmacovigilance are yet to be realized.

1.1 Semantic Web Technologies and Linked Data

The Semantic Web was conceived with the vision that a decentralized, distributed and heterogeneous data space, extending over the traditional Web, can reveal hidden associations that are not directly observable.[12] Semantic Web technologies and linked data principles enable the representation, linking and querying of data and knowledge on the Web.[13] Semantic Web technologies include the W3C standards Resource Description Framework (RDF) and the SPARQL graph query language.[14,15] Due to the challenges of integrative bioinformatics, biomedical researchers have been the earliest adopters of Semantic Web technologies and linked data principles to create the Life Sciences Linked Open Data (LSLOD) cloud.[16] Biomedical data and knowledge sources are converted to graphs using the RDF model. SPARQL can use specific graph patterns to query these RDF graphs. Several different efforts publish and link biomedical data and knowledge in the LSLOD cloud (e.g., Bio2RDF[17]). Several sources that may be relevant to systems pharmacology, such as PharmGKB and DrugBank,[18,19] are made available through the LSLOD cloud.

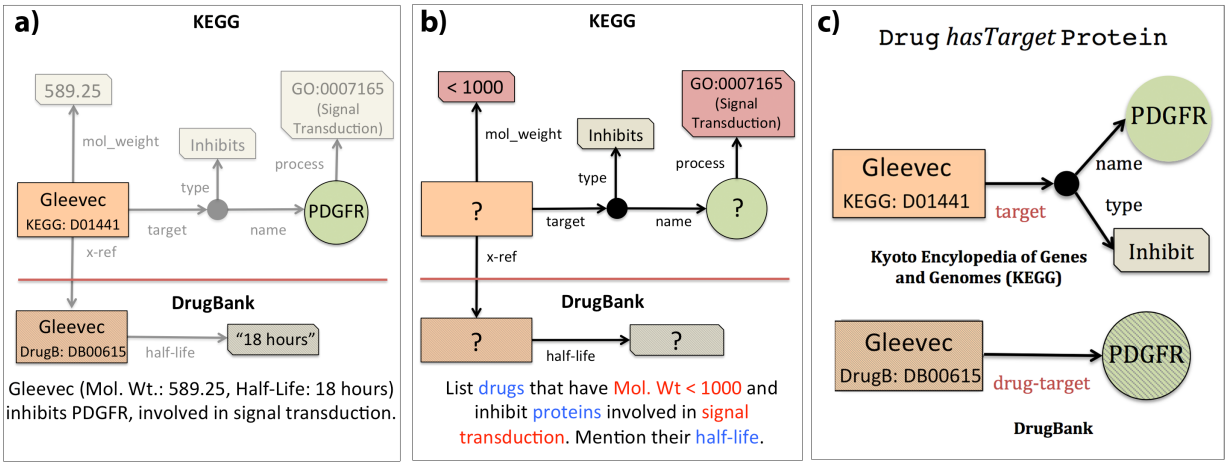

Figure 1: Semantic Web Technologies: **a)** Data from two different sources – KEGG and DrugBank are represented as RDF graphs, and similar entities (`Gleevec`) are linked together using explicit *x-ref* attributes. **b)** The SPARQL Graph Query Language can be used to query these RDF graphs and retrieve information from multiple sources. The opaque nodes in **a)** highlight the information retrieved by the SPARQL query in **b)**. **c)** Different graph patterns may be used by different sources to capture the same relation type (`Drug` *hasTarget* `Protein`), hence a pattern-based query federation is necessary to retrieve the relations from two sources.

An example of an RDF graph that represents the following information — *"Gleevec (Mol. Wt.: 589.25, Half-Life: 18 hours) inhibits PDGFR (platelet-derived growth factor receptor), involved in signal transduction"* — is shown in **Figure 1a**. Here, similar entities (e.g. `Gleevec`) in two different sources, DrugBank[19] and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[11], are linked together using explicit *x-ref* (cross-reference) attributes. Attributes and relations stored in these data sources are represented as nodes and edges in the RDF graph. Moreover, as shown in **Figure 1b**, the following query — *"List drugs that have Mol. Wt $<$ 1000 and inhibit proteins involved in signal transduction. Mention their half-life"* — can be executed using SPARQL. The graph expression patterns are derived from the RDF schema that is used to structure the graphs, as well as the *x-ref* attributes.

It should be noted that these biomedical RDF graphs may be exposed through isolated SPARQL endpoints on the web. Querying multiple isolated SPARQL endpoints simultaneously over the web requires a scalable SPARQL **query federation** method.[20] Different graph patterns may be used to represent the same relation type. In **Figure 1c**, the relation `Drug` *hasTarget* `Protein` is represented using different labels (*drug-target* and *target*) and different graph patterns in DrugBank and KEGG respectively. In the latter case, as KEGG is a pathway data source, the RDF graph also captures the type and provenance of the interaction between the `Drug` and `Protein`. Query Federation methods can transform a given query to source-specific queries and retrieve information from two or more sources simultaneously.

In our previous research,[21] we had developed such a query federation architecture, termed PhLeGrA – Linked Graph Analytics in Pharmacology, over the LSLOD cloud. PhLeGrA uses prior knowledge on such graph patterns to generate a systems pharmacology network by retrieving data from four different biomedical sources. As there is minimal overlap between different sources for drugs and drug–protein relations, we had also demonstrated how query federation over the LSLOD cloud can help systems pharmacology approaches.

In this research, we extend the PhLeGrA architecture, with an improved inference module to detect drug–drug interactions and adverse drug reactions. The module will also assign confidence scores to all possible underlying biological mechanisms for the DDIs. The key contributions of this research can be outlined as follows:

1. We propose and implement a graph analytics method, inspired from the Apriori algorithm[22] for association rule mining, to identify frequent substructures in our systems network, as derived from mining FAERS reports.
2. We compare our method with two baseline methods in pharmacovigilance – the Gamma Poisson Shrinker (GPS) method and the Bayesian Confidence Propagation Neural Network (BCPNN) over three different validation sets.
3. We discuss briefly, the insights obtained from our method on few drug–ADR associations as well as discuss the advantages of using Semantic Web Technologies for mechanism-based pharmacovigilance.

All the results described in this paper, as well as prior research on the PhLeGrA platform, are available online at `http://onto-apps.stanford.edu/phlegra/`.

2 Related Work

Several methods have been developed to predict DDIs, or predict ADRs that manifest due to concomitant intake of multiple drugs, by mining spontaneous reporting systems such as FAERS or electronic medical records. Harpaz et al.[7,22] used the Apriori algorithm to mine the FAERS reports and generate statistically significant association rules between multiple drugs and ADRs (e.g. $Aspirin + Warfarin \rightarrow Bleeding$). Iyer et al.[8] used electronic health records and generated patient timelines of drug and ADR mentions in the records. Using adjusted disproportionality ratios to identify significant drug–drug–event associations, and a manually-curated gold standard of such associations from Drugs.com and MediSpan, they demonstrated that their approach can be used to complement FAERS mining for pharmacovigilance. Bayesian approaches such as the Multi-Item Gamma Poisson Shrinkage (MGPS) algorithm[23] and the Bayesian Confidence Propagation Neural Network (BCPNN)[24], as well as approaches using existing knowledge on drug and ADR similarities[25], have recently been proposed to deal with reporting bias and confounding factors, observed in Spontaneous Reporting Systems. The performance of these methods are compared by Harpaz, et al.[26] However, these methods fail to demonstrate the possible underlying molecular mechanisms behind these associations.

Systems pharmacology methods[9,27] have also been explored in the context of drug–ADR association discovery or drug repurposing (use of existing drugs to treat new conditions). These methods generally combine databases and knowledge bases, to generate a systems network, manually without the use of Semantic Web technologies. *CauseNet*[28] combines four biomedical sources into a k-partite network for generating new drug repurposing hypotheses. Berger, et al.[29] integrated diseases with the human protein–protein interaction network to understand the systems pharmacology underlying specific forms of drug-induced arrhythmias. While these approaches are similar to our research, our method retrieves data and knowledge from the LSLOD cloud and can generate such systems networks more easily[21].

The LSLOD cloud has been utilized to predict new DDIs recently. *Tiresias* processes drug-related data and knowledge and predicts new DDIs using large-scale similarity matching[30]. Most approaches consider binary drug pairs and not multiple drug interactions[31], they ignore the underlying molecular mechanisms, and they may not associate the adverse drug reactions with the DDIs[32]. Noor et al.[33] constructed a mechanism-based DDI knowledge warehouse by integrating knowledge from multiple sources in the LSLOD cloud at the pharmacokinetic, pharmacodynamic, and pathway interaction level, and developed an inference engine to generate mechanistic explanations for DDIs. However, this method does not rank the mechanistic explanations, is not implemented for pharmacovigilance, and due to the knowledge warehouse, updates in the underlying sources are not captured instantaneously.

3 Materials and Methods

3.1 PhLeGrA network generation

In this section, we summarize the query federation method to extract a k-partite network from multiple, heterogeneous biomedical data sources available through the Life Sciences Linked Open Data Cloud (LSLOD). The method is described in more detail in Kamdar, et al.[21]. We integrate four different data sources that are published as SPARQL endpoints by the Bio2RDF project[17] (Version 4) in the LSLOD cloud – DrugBank[19], PharmGKB[18], Kyoto Encyclopedia for Genes and Genomes (KEGG)[11] and Comparative Toxicogenomics Database (CTD)[34]. These four sources contain data and knowledge on drugs, proteins, pathways, phenotypes and their inter-connections (e.g. drug–protein target relations) and have been used in several pharmacological methods previously.

We use a pattern-based query federation method[20,21] to query the SPARQL endpoints of these sources simultaneously to generate the k-partite systems pharmacology network. Specifically, we retrieve four different types of entities — **(E1)** `Drug`, **(E2)** `Protein`, **(E3)** `Pathway`, and **(E4)** `Phenotype` (adverse drug reaction). We also retrieve five different types of biological relations — **(R1)** `Drug` *hasTarget* `Protein`, **(R2)** `Drug` *hasEnzyme* `Protein`, **(R3)** `Drug` *hasTransporter* `Protein`, **(R4)** `Protein` *isPresentIn* `Pathway`, and **(R5)** `Pathway` *isImplicatedIn* `Phenotype`. The SPARQL graph patterns used to retrieve the entities and relations from the sources are listed at `http://onto-apps.stanford.edu/phlegra/about`.

The entities and relations, retrieved from the LSLOD cloud, form a k-partite network — a network whose nodes can be partitioned into k different independent sets ($k = 4$). We decided on these types of entities and relations to capture the following underlying mechanisms behind drug–drug interactions: *a)* one drug may inhibit the enzymes that metabolize a second drug to its inactive or active state, *c)* one drug may inhibit the transporters that decrease the absorption or elimination of a second drug, *c)* two drugs may target the same protein, leading to varying effects of both drugs, or *d)* two drugs may target proteins in the same pathway leading to varying effects of both drugs. Hence, here we consider transporters and enzymes to be considered as specialized proteins.

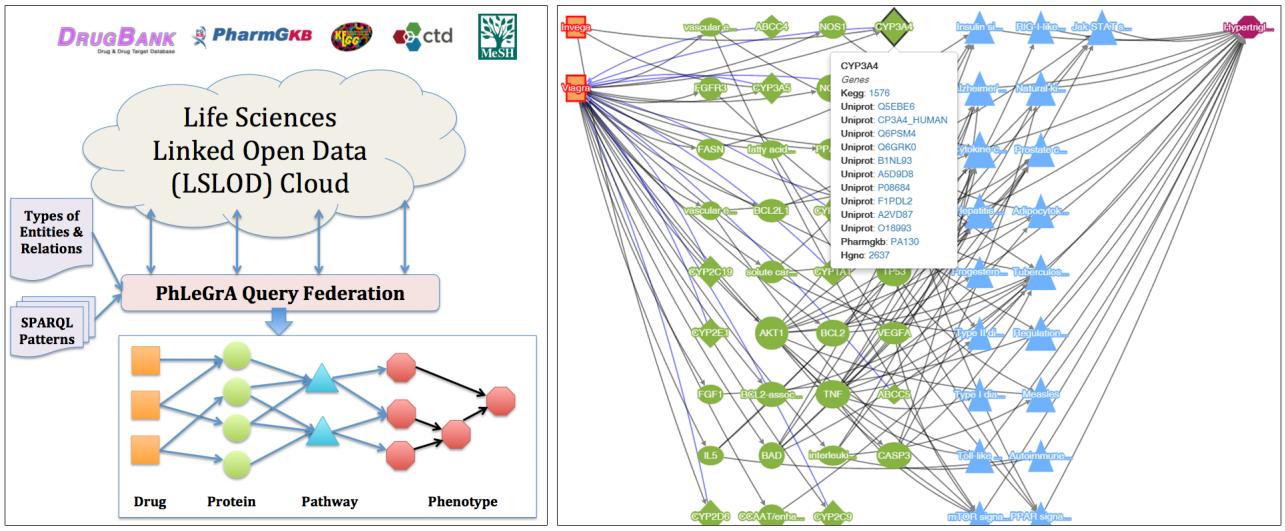

Figure 2: PhLeGrA network generation. **a)** PhLeGrA query federation method uses the type of entities and relations, as well as prior knowledge on SPARQL graph patterns to query four sources (DrugBank, KEGG, PharmGKB and CTD) in the LSLOD cloud to create a k-partite network composed of drugs, proteins, pathways and phenotypes. The phenotypes are further arranged using the MESH hierarchy tree. **b)** A visualization of a small portion of the network, with the Drugs `Invega` and `Viagra`, Enzyme `CYP3A4` and Phenotype `Hypertriglyceridemia` highlighted.

To reconcile similar entities in different sources (e.g. drugs present in KEGG and DrugBank referenced using different identifiers), we use the *x-ref* attributes provided by the Bio2RDF project. We reconcile entities to a uniform identifier

nomenclature scheme using existing terminologies – ATC (Anatomical Therapeutic Chemical Classification[35]) for `Drug`, HGNC (Hugo Gene Nomenclature Committee[10]) for `Protein`, KEGG for `Pathway`, and MESH (Medical Subject Headings[36]) for `Phenotype`. We further organize the `Phenotype` MESH identifiers into a hierarchy, using the MESH hierarchy. As Spontaneous Reporting Systems (e.g. FAERS) collect patient reports in which the adverse reactions may be specified at different levels of abstraction, this step helps in aggregation of reports on higher abstract terms. For example, there may be patient reports on both `Anaphylaxis` and `Hypersensitivity, Immediate`, where the former `Phenotype` term is a subclass of the latter term. Hence the k-partite network is coupled with the MESH `Phenotype` hierarchy. A visualization of such a network is shown in **Figure 2b**.

3.2 FDA Adverse Event Reporting System

Spontaneous reporting systems are the primary means to conduct post-marketing surveillance of drug products to detect ADRs that were not determined during clinical trials. The US Food and Drug Administration (FDA) collects reports on the adverse drug reactions observed in patients subjected to multiple drugs. The FDA Adverse Event Reporting System[37] (FAERS), a public data portal, publishes these reports after the anonymization of the patient data. We downloaded the FAERS datasets for three years from January 2013 to December 2015. Each dataset is composed of several safety reports. Among many features, each safety report indicates the set of ADRs observed in a patient (e.g., heart attack), and the set of drugs administered to the patient (e.g., Sildenafil). The steps taken to process and align the FAERS records with the `Drug` and `Phenotype` nodes in our k-partite network are described previously[21].

3.3 Frequent Substructure Mining

We extend the method proposed by Harpaz, et al.[7,22] for statistical mining in FAERS datasets. This method is inspired from the Apriori algorithm to mine association rules (e.g. $\{Drug\}_n \rightarrow ADR$) in large databases, in an unsupervised, computationally tractable way. The Apriori algorithm prunes the search space of associations, such that if a certain combination of drugs and ADRs is infrequent, then any larger combination that builds upon the smaller infrequent one, will also be infrequent. Certain thresholds can be decided for ignoring these combinations.

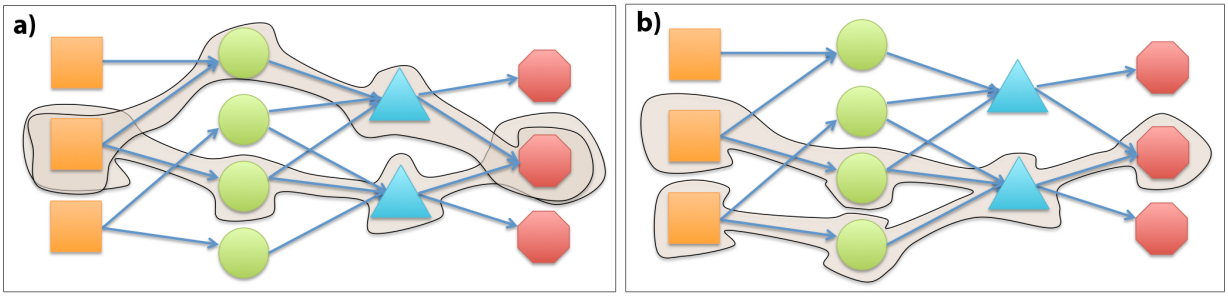

Figure 3: Cartoon representations of the propagation of the FAERS reports along the networks. **a)** All nodes on the shortest path, linking a Drug to an ADR are annotated with the FAERS report that mentions these terms. **b)** We only consider substructures during the association rule mining of the form $\{Drug\}_n \rightarrow ADR$, such that the paths linking the different Drugs to the ADR have either a Protein or a Pathway node in the intersection set.

The Apriori algorithm has also been modified to mine frequent substructures in graphs.[38] We have used this implementation of the Apriori algorithm to work on k-partite networks. Specifically, we can determine the set of FAERS reports that contain any specific (drug, ADR) pair. As shown in **Figure 3a**, we propagate the set of reports along all the possible shortest, directed, paths that connect the corresponding `Drug` node to the `Phenotype` node in the k-partite network. We decided to use only the shortest paths to make the method computationally tractable. Hence, each node in the k-partite network is annotated with the set of reports it may be associated with. We are unaware of the underlying biological mechanisms at this point, so all implicit associations are equally probable.

Generally, Apriori-based methods compute the **Support** and **Confidence** statistics for an association rule. Suppose,

$S(A)$ indicates the number of reports that describe the items in itemset A (e.g. set of drugs). Then the support for an association rule is simply $S(A \rightarrow B) = S(A \cup B)$ (i.e. number of reports where the items in itemsets A and B cooccur). The confidence of an association rule can be described as $C(A \rightarrow B) = S(A \cup B)/S(A)$ (i.e. the conditional probability for observing items in itemset B, given items in itemset A). The space of all possible association rules is pruned by selecting only those itemsets that exhibit a minimum value for the support statistic. In our method, as we have four different types of nodes, the association rules are generated by observing a minimum support at each step, where the direction of adding new itemsets is strictly `Drug` $\rightarrow$ `Protein` $\rightarrow$ `Pathway` $\rightarrow$ `Phenotype`. Nodes and edges that do not exhibit a support statistic that exceeds a given threshold are automatically pruned from the k-partite network. To further reduce the number of computations, we only consider $Drug \rightarrow ADR$ associations such that they have some direct path in the k-partite network, and we only consider $\{Drug\}_2 \rightarrow ADR$ associations such that the paths that link the two drugs to the ADR have a common intersection point, either at the `Protein` or the `Pathway` node in the network. This method and our optimizations are visually explained in **Figure 3b**. It should be noted that this method can be extended to include multi-drug interactions ($\{Drug\}_n \rightarrow ADR$).

The confidence statistic is computed as $C(\texttt{Drug} \cup \texttt{Protein} \cup \texttt{Pathway} \rightarrow \texttt{Phenotype})$, and is used to rank the different substructures (i.e. different underlying mechanisms), that lead to the manifestation of the ADR given the set of drugs. However, due to the reporting bias in FAERS, this statistic in itself is not sufficient to actually determine if there is any association between the drugs and the ADR, as indicated by Harpaz, et al.[7] Hence, we also compute a **Network-based Relative Reporting Ratio (RRR)** statistic, that considers the $Actual/Expected$ ratio at each path in the k-partite network. Hence, each possible substructure has an RRR statistic. RRR is defined as the ratio between an association rule's observed frequency to a baseline expected frequency under the assumption of independence.

$$RRR = \frac{N \times S(A \cup B)}{S(A) \times S(B)}$$

Here $A = \texttt{Drug} \cup \texttt{Protein} \cup \texttt{Pathway}$, N is the total number of FAERS records, and $B = \texttt{Phenotype}$. The median value of the RRR, computed for relevant substructures, is used as the statistic to compare our method against other baseline methods. To summarize, **Support** statistic is used to prune the k-partite network, **Network-based Relative Reporting Ratio** is used to determine whether an association between a set of drugs and an ADR exists, and **Confidence** statistic is used to rank the different underlying mechanisms behind $\{Drug\}_n \rightarrow ADR$ association.

3.4 Method Evaluation

We collected three different datasets that consist of manually-curated positive and negative drug–adverse reaction associations. These datasets have been used to validate DDI prediction methods previously. The Observational Medical Outcomes Partnership (OMOP[39]) dataset and the European "Exploring and Understanding Adverse Drug Reactions" project (EU-ADR[40]) dataset consists of single drug–ADR associations. The dataset described in Iyer, et al.[8] consists of drug–drug–ADR associations retrieved from Drugs.com and MediSpan.

Table 1: The coverage of different validation datasets used in this study.

Dataset	Unique Drugs	Unique ADRs	Positive associations	Negative associations
OMOP	155	4	137	158
EU-ADR	59	9	44	39
Iyer et al.[8]	252	9	315	288

Some statistics for these datasets, in terms of positive and negative associations, as well as coverage of drugs and ADRs are shown in **Table 1**. All the three validation sets were transformed, such that the drugs were referenced using ATC identifiers and ADRs were referenced using MESH identifiers. Some common ADRs across the three datasets include – Gastrointestinal Hemorrhage, Hyperkalemia, Acute Kidney Injury and Drug-induced Liver Injury. Using these validation sets, we compare our method with two baseline methods — the Gamma Poisson Shrinkage (GPS) method and the Bayesian Confidence Propagation Neural Network (BCPNN) method. We used the R package for PharmacoVigilance Signal Detection (PhViD[1]) for the baseline methods.

[1] `https://cran.r-project.org/web/packages/PhViD/PhViD.pdf`

4 Results

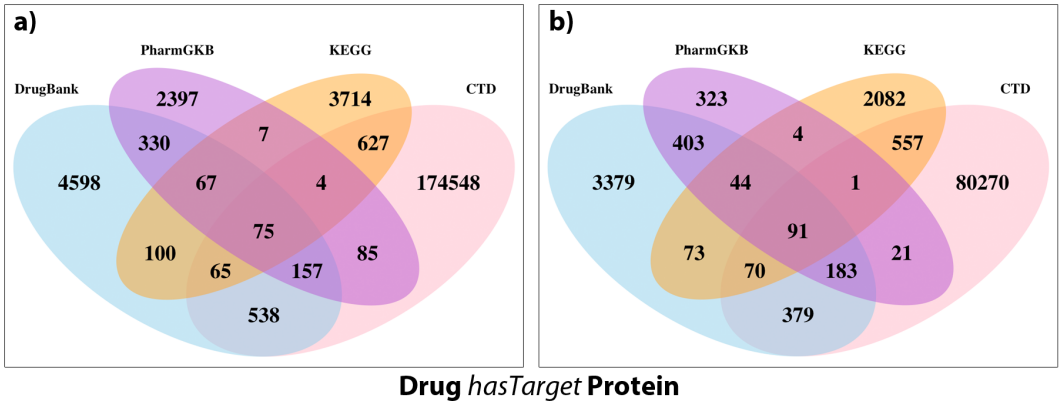

Figure 4: The source distribution of the `Drug` *hasTarget* `Protein` before (**a**) and after (**b**) pruning the k-partite network on the basis of a minimum **support** value for each node and edge. It can be seen that the drug–protein relations from the CTD source are reduced by more than half, the original number.

The k-partite network that was generated from the PhLeGrA query federation method, consisted of 2,759 drugs (**E1**), 3,890 phenotypes (**E4**) organized using the MESH hierarchy, 19,903 genes (**E2**) and 301 pathways (**E3**). The network also consisted of 249,001 drug–target relations (**R1**), 2,062 drug–enzyme relations (**R2**), 919 drug–transporter relations (**R3**), 25,480 protein–pathway relations (**R4**) and 46,300 pathway–phenotype relations (**R3**). Individual statistics for the different entities and relations extracted from each source were presented previously by Kamdar, et al.[21]. We used ≈ 3 million FAERS reports for the frequent substructure mining method demonstrated in this research.

The FAERS reports were propagated along our k-partite network, with each node in the network annotated with the set of nodes that it may be associated with. We use a **Support** threshold of 200 to filter out nodes and edges in the k-partite network. After applying the support threshold, we were able to decrease our k-partite network to include only 7,217 nodes and 89,451 edges. Moreover, as seen from the **Figure 4**, the number of entities and relations of a specific type (e.g. `Drug` *hasTarget* `Protein`) is reduced drastically for a particular source (e.g. CTD). It can be argued that our method can remove spurious relations in the k-partite network that may not be relevant, or may be incorrect.

We compute the Network-based RRR statistic for a given association between a set of drugs and an ADR, given all possible substructures. We compare this statistic with the GPS and BCPNN statistic over the three validation datasets. The Receiver-Operator Characteristic curve for each validation dataset is shown in **Figure 5**. It can be seen that the area under the curve (AUROC) statistic for each of the three validation sets is almost similar to the baseline methods, and the value is around 0.7–0.8. The AUROC statistic actually exceeds by 0.01–0.02 over the baseline methods for the OMOP and the EU-ADR validation sets. Moreover, as observed by Iyer et al.[8], using event-specific thresholds on the statistic can actually generate higher AUROCs for certain ADRs. This is observed in **Figure 5d** where we obtain an AUROC of almost 0.94 for rhabdomyolysis. Finally, we would like to note that for higher values of Specificity, our sensitivity is considerably less than existing baseline methods. This may be due to the reporting bias in FAERS, which is not tackled in this research. However, using GPS-adjusted Expected values in the Network-based RRR statistic actually alleviated this issue (plot not shown). On a closer inspection of the false positive associations, we observed that these associations have more connecting paths in the k-partite network when compared to the true negative associations. Similarly, the false negative associations have fewer connecting paths in the k-partite network when compared to the true positive associations. Using a t-test, we found that these comparison findings were statistically significant ($p < 0.05$). Hence, the topology of the network has an impact on the association discovery method.

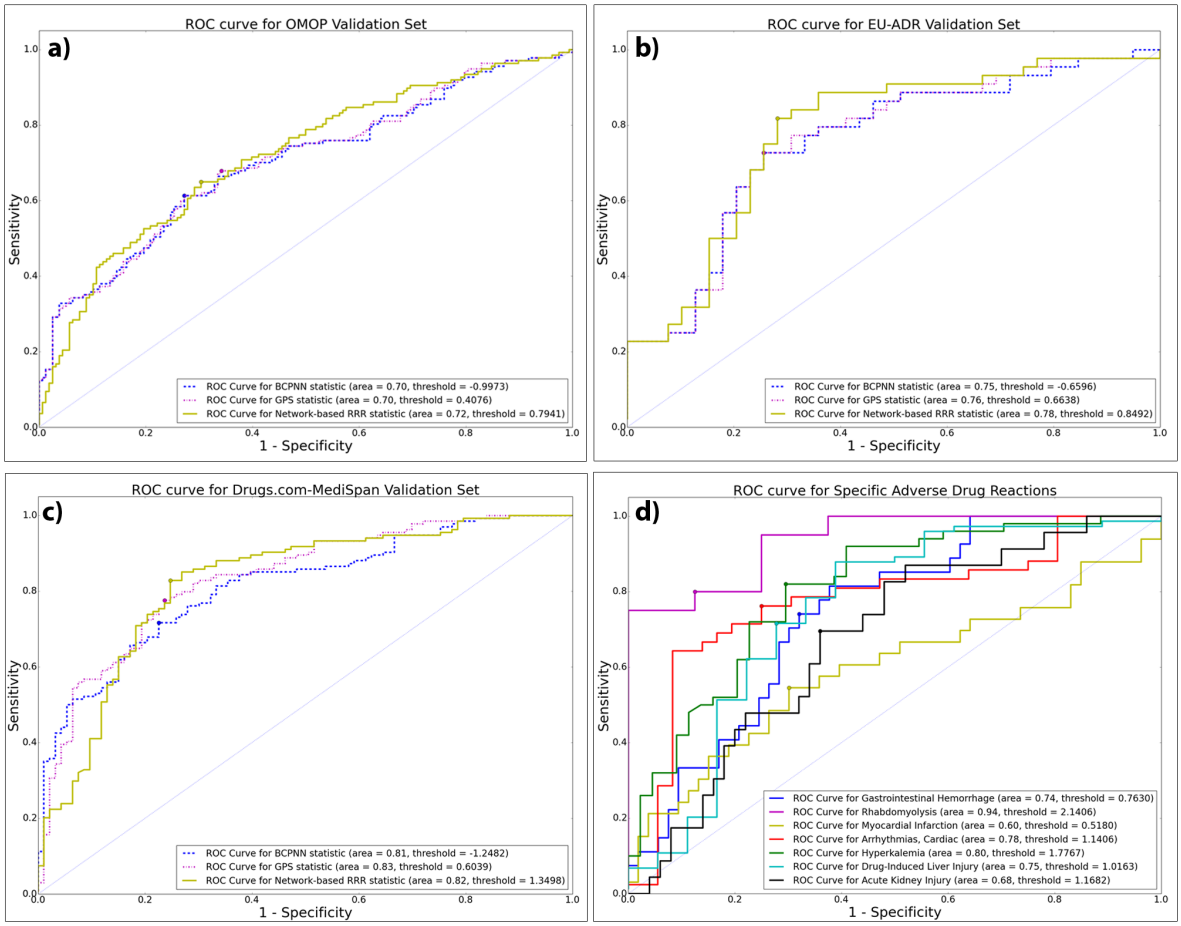

Figure 5: The first three plots include the ROC (Receiver-Operator Characteristic) curves for the three validation datasets used in our study – **a)** OMOP, **b)** EU-ADR and **c)** Drugs.com-MediSpan corpus curated by Iyer, et al.[8] The predictive power of our Network-based RRR method is represented using the solid yellow line, whereas the dotted magenta and the dotted blue line indicate the predictive power of the GPS and BCPNN methods respectively. In **d)**, we demonstrate ADR-specific predictions using the Network-based RRR method.

5 Discussion

Using the **confidence** statistic that is also computed by our method, we were able to observe some interesting and some known substructures. For example, while it is known that simvastatin may interact with the CYP3A4 inhibitor itraconazole to cause rhabdomyolysis, the corresponding substructure was observed to have a high confidence value in our analysis. Moreover, we found that the drug paliperidone, which is another CYP3A4 inhibitor and is used as a treatment in bipolar disorder, may interact with several other drugs to cause hypertriglyceridemia, hyperprolactinemia and gynecomastia. However, these findings need to be validated by a domain expert in the future. We plan to incorporate this method in the PhLeGrA visualization browser[2], such that different substructures can be highlighted and ranked with the support and confidence statistics and can provide a better understanding to the domain user.

Currently, our method is limited to substructure discovery from only those parts in the networks where there exist some edges that connect the different entity types (e.g. two drugs may interact with the same proteins, but this knowledge is derived from an existing source). This assumption was used to allow our Apriori algorithm to terminate under a reasonable runtime. However, there may exist drug–protein relations that are not known currently, or may not be stored

[2] `http://onto-apps.stanford.edu/phlegra/`

in the biomedical sources that were integrated to generate the k-partite network. We will argue that the benefits of the Semantic Web technologies and Linked Data principles allows us to incorporate multiple data sources in the k-partite network, whereas the initial thresholding using the **Support** statistic can allow us to filter irrelevant edges and nodes in the network. This was observed for the Comparative Toxicogenomics Database that incorporated a huge proportion of noisy edges between drugs and protein targets that were not actual relations.[21]

As presented here, pattern-based query federation[21] can bring together pharmacological knowledge existing in isolated, heterogeneous sources without being concerned about the underlying semantics and schema differences. This advantage, when coupled with the network-based Apriori method for association rule mining, can facilitate domain users to obtain mechanistic explanations behind detected DDIs and ADRs, as well as generate new knowledge on underlying biological mechanisms (frequently observed substructures). Such systems pharmacology networks, as previously described[21] are easier to generate using Semantic Web technologies and query federation methods.

6 Conclusion

In this research, we have demonstrated a method for mechanism-based pharmacovigilance from Spontaneous Reporting Systems, such as the FAERS datasets. While our method has equivalent, if not better, performance compared to existing state-of-the-art methods in pharmacovigilance, it can also be used to provide a mechanistic understanding behind the drug–drug interactions and the adverse reactions that manifest on the account of those DDIs. Moreover, it can enable biomedical researchers to obtain newer knowledge on molecular mechanisms that may be relevant in pharmacovigilance, or may be spurious in a particular database. We use Semantic Web technologies to easily generate a systems pharmacology network, and to the best of our knowledge this is the first approach to provide explanations of underlying biological mechanisms using a ranking scheme.

Acknowledgments

The authors acknowledge Rainer Winnenburg, Erik van Mulligen and Juan Banda for providing the OMOP, EU-ADR and Drugs.com-MediSpan datasets respectively. The authors also acknowledge Michel Dumontier for his help using Bio2RDF linked data. This work is supported by Grant HG004028 from the US National Institutes of Health.

References

1. Reamer L Bushardt et al. Polypharmacy: misleading, but manageable. *Clinical interventions in aging*, 3(2):383, 2008.
2. Jia Jia et al. Mechanisms of drug combinations: interaction and network perspectives. *Nature reviews Drug discovery*, 8(2):111–128, 2009.
3. Johanna Strandell et al. Drug–drug interactions–a preventable patient safety issue? *British journal of clinical pharmacology*, 65(1):144–146, 2008.
4. Dorothy Bonn. Adverse drug reactions remain a major cause of death. *The Lancet*, 351(9110):1183, 1998.
5. Jason Lazarou et al. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205, 1998.
6. Frank R Ernst et al. Drug-related morbidity and mortality: updating the cost-of-illness model. *Journal of the American Pharmaceutical Association (1996)*, 41(2):192–199, 2001.
7. Rave Harpaz et al. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC bioinformatics*, 11(9):S7, 2010.
8. Srinivasan V Iyer et al. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association*, 21(2):353–362, 2014.
9. Jane PF Bai et al. Systems pharmacology to predict drug toxicity: integration across levels of biological organization*. *Annual review of pharmacology and toxicology*, 53:451–473, 2013.
10. Sue Povey et al. The HUGO gene nomenclature committee (HGNC). *Human genetics*, 109(6):678–680, 2001.
11. Minoru Kanehisa et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 28(1):27–30, 2000.
12. Tim Berners-Lee et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

13. Christian Bizer et al. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.
14. Graham Klyne et al. Resource description framework (RDF): Concepts and abstract syntax. 2006.
15. Eric PrudHommeaux et al. SPARQL query language for RDF. *W3C recommendation*, 15, 2008.
16. Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
17. Alison Callahan et al. Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In *The semantic web: semantics and big data*, pages 200–212. Springer, 2013.
18. Micheal Hewett et al. PharmGKB: the pharmacogenetics knowledge base. *Nucleic acids research*, 30(1):163–165, 2002.
19. David S Wishart et al. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672, 2006.
20. Maulik R Kamdar et al. ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. *Journal of biomedical informatics*, 47:112–130, 2014.
21. Maulik R. Kamdar et al. PhLeGrA: Graph analytics in pharmacology over the web of life sciences linked open data. In *Proceedings of the 26th World Wide Web Conference, WWW 2017, Perth*, 2017.
22. Rave Harpaz et al. Statistical mining of potential drug interaction adverse effects in fdas spontaneous reporting system. In *AMIA Annu Symp Proc*, volume 2010, pages 281–285, 2010.
23. Ana Szarfman et al. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fdas spontaneous reports database. *Drug Safety*, 25(6):381–392, 2002.
24. Andrew Bate. Bayesian confidence propagation neural network. *Drug safety*, 30(7):623–625, 2007.
25. Nicholas P Tatonetti et al. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.
26. Rave Harpaz et al. Performance of pharmacovigilance signal-detection algorithms for the fda adverse event reporting system. *Clinical Pharmacology & Therapeutics*, 93(6):539–546, 2013.
27. Seth I Berger et al. Role of systems pharmacology in understanding drug adverse events. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(2):129–135, 2011.
28. Jiao Li et al. Pathway-based drug repositioning using causal inference. *BMC bioinformatics*, 14(16):1, 2013.
29. Seth I Berger et al. Systems pharmacology of arrhythmias. *Science signaling*, 3(118):ra30, 2010.
30. Achille Fokoue et al. Predicting drug-drug interactions through large-scale similarity-based link prediction. In *International Semantic Web Conference*, pages 774–789. Springer, 2016.
31. Juan M Banda et al. Feasibility of prioritizing Drug–Drug-Event Associations found in Electronic Health Records. *Drug safety*, 39(1):45–57, 2016.
32. Juan M Banda et al. Provenance-centered Dataset of Drug-Drug Interactions. In *The Semantic Web-ISWC 2015*, pages 293–300. Springer, 2015.
33. Adeeb Noor et al. Drug-drug interaction discovery and demystification using semantic web technologies. *Journal of the American Medical Informatics Association*, page ocw128, 2016.
34. Allan Peter Davis et al. The comparative toxicogenomics database: update 2013. *Nucleic acids research*, 41(D1):D1104–D1114, 2013.
35. World Health Organization. Anatomical therapeutic chemical (ATC) classification index with defined daily doses (DDDs). *Oslo: WHO Collaborating Centre for Drug Statistics Methodology*, 2000.
36. Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.
37. Food and US Drug Administration. *FDA Adverse Event Reporting System*, 2013. (March 01, 2016).
38. Akihiro Inokuchi et al. An apriori-based algorithm for mining frequent substructures from graph data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 13–23. Springer, 2000.
39. Paul E Stang et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Annals of internal medicine*, 153(9):600–606, 2010.
40. Erik M Van Mulligen et al. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884, 2012.