2018 IEEE 34th International Conference on Data Engineering

MeDIAR: Multi-Drug Adverse Reactions Analytics

Xiao Qin*[1,2], Tabassum Kakar*[1,2], Susmitha Wunnava[1], Brian McCarthy[1], Andrew Schade[1], Huy Quoc Tran[1], Brian Zylich[1], Elke A. Rundensteiner[1], Lane Harrison[1], Sanjay K. Sahoo[2], and Suranjan De[2]

[1]Department of Computer Science, Worcester Polytechnic Institute
[2]Center for Drug Evaluation and Research, U.S. Food and Drug Administration
[1]{xqin,tkakar,swunnava,bmccarthy,alschade,hqtran,bzylich,rundenst,lharrison}@wpi.edu
[2]{xiao.qin, tabassum.kakar, sanjay.sahoo, suranjan.de}@fda.hhs.gov

***Abstract*—Adverse drug reactions (ADRs) caused by drug-drug interactions (DDI) are a major cause of morbidity and mortality worldwide. There is a growing need for computing-supported methods that facilitate the automated signaling of DDI related ADRs (DIARs) that otherwise would remain undiscovered in millions of ADR reports. In this demonstration, we showcase our MeDIAR technology – an end-to-end DIAR signal generation, exploration and validation solution for pharmaceutical regulatory agencies to detect true DIAR signals from a drug surveillance database. MeDIAR's innovations include an efficient rule-driven learning algorithm for deriving DIAR signals from ADR reports, an innovative scoring methodology based on the proposed contextual association cluster model to rank the generated signals by their importance. Further, these ranked signals are augmented with meta information such as their *significance level* and their *severity*, along with *links* to their supporting ADR reports. Lastly, MeDIAR features an interactive visual analytics interface to support drug safety evaluators in reviewing and discovering unknown severe DIARs.**

I. Introduction

A. Motivation

Adverse Drug Reactions (ADRs), one of the leading causes of death, correspond to unwanted and often extremely dangerous effects caused by the administration of drugs. *Polypharmacy*, the use of multiple drugs taken simultaneously to treat medical conditions, can lead to ADRs due to drug-drug interactions (DDI). For example, *Aspirin* taken together with *Warfarin* may lead to excessive bleeding [1]. This interaction of the two seemingly harmless drugs if taken in combination can be life threatening. It is critical that such DDI related ADRs (DIARs) are detected early with minimum patient exposure to avoid further harmful incidents. Unfortunately, identifying all DIARs in clinical trails is prohibitively expensive and impossible in practice – as it would require exhaustive testing of every possible drug combination on human subjects.

For early detection of DIARs not captured during clinical trials, Spontaneous Reporting Systems (SRS) have been adopted in the US [2] and elsewhere [3] to collect information on adverse events related to drugs reported by patients, health care professionals and drug manufacturers. The massive databases collected by these surveillance programs represent a valuable resource to tap into for DIAR discovery.

*Authors contribute equally to this work.

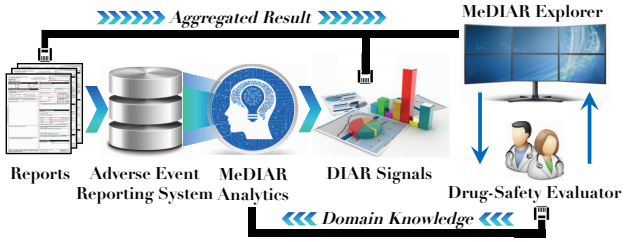

Fig. 1. The MeDIAR system

In 2016, U.S. Food and Drug Administration (FDA) received approximately 1.7 million ADR reports – with this number growing exponentially. Discovering DIAR signals by reading and analyzing all these reports one by one is difficult for drug safety evaluators. Instead, computational methods, especially data mining techniques, offer the promise to revolutionize the process by discovering emerging DIAR signals from these data repositories. These mined signals, along with the reports from which these signals were derived, could then be recommended to the drug safety evaluator as potential DIAR hypothesis for further investigation and validation.

B. Limitations of State-of-the-Art

To tackle this problem of discovering interesting relationships among variables in the data, called DIAR signaling, Association Rule Learning (ARL) has been identified as a suitable data mining method. Wei et al. [4] adopted ARL to find vaccine DIARs in the FDA Vaccine Adverse Event Reporting System (VAERS); while Harpaz et al. [5] applied ARL with *lift* [6] measure to signal DIARs in FDA Adverse Event Reporting System (FAERS) [2]. However, directly applying traditional ARL algorithms to detect DIAR signals produces a huge number of associations [7] – most of which are either previously known, redundant, or even misleading. Thus there is a need to develop a mechanism that effectively mines and ranks non-spurious associations based on their likelihood of being a real DIAR.

C. Research Challenges

To develop an end-to-end DIAR analytics solution using ARL for drug regulatory agencies, the following research challenges must be addressed:

2375-026X/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDE.2018.00177

IEEE computer society

Modeling DIAR signals and misleading signals. Without pre-established dependency constraints among items appearing in each record in the database, existing ARL algorithms consider every possible combination of items that appears in a transaction as an *itemset*. This results in a huge amount of *redundant* [8] even *misleading* [7] associations in the context of signaling DIAR from ADR reports. The non-spurious association must be clearly defined and efficiently generated.

Scoring and ranking DIAR signals. Given n distinct drugs and m unique ADRs that appear in a report set, ARL generates up to $\mathcal{O}(2^{n+m})$ possible DIAR signals – clearly impossible for a drug-safety evaluator to sift through. Although, the off-the-shelf association measures such as *support* and *confidence* [6] could be used to rank the signals, this amounts to counting the co-occurrence of the modeled drugs and ADRs. However, these simple measures are not able to capture true interactions among drugs. Therefore, appropriate DIAR metrics must be developed to rank the generated signals.

Exploring and interpreting augmented signals. First, we need to model the different types of drug and ADR interactions. Further, external information about the generated associations would need to be extracted and integrated into the DIAR model to enable the safety evaluators to correctly interpret the results. One such meta data concerns if the generated DIAR signals were already previously known or if there is a sudden spike in their occurrence. To facilitate evaluators in their exploration of this complex information, intuitive visual displays and interactions that seamlessly bridge different views of signals augmented with metadata must be designed.

D. The MeDIAR Approach

MeDIAR generates non-spurious DIAR signals from the drug surveillance database based on our observation that a small set of associations – the *closed* Drug-ADR associations [7] out of the huge number of associations produced by traditional ARL techniques is sufficient to assist the evaluator to discover all potential DIARs. By analyzing the factors that influence the significance of the discovered associations, we propose a new association measure called *contrast* [7] to quantify how likely the ADRs that appear in a signal are triggered by the interaction of the modeled drugs instead of the individual ones. MeDIAR augments the generated signals with meta data extracted from medical web repositories to improve interpretability. MeDIAR is equipped with interactive visualizations to support the safety evaluator in her exploration on the *macroscopic* and *microscopic* level.

Our contributions include:

- We design and implement an end-to-end DIAR signaling solution that facilitates the drug safety evaluator to identify emerging severe DIARs.
- We adopt and adapt ARL for DIAR modeling and signal generation. We propose a pruning strategy to remove *spurious* associations while keeping relevant ones.
- We propose the *contextual association cluster* model and the *contrast* measure to evaluate the interestingness of the associations and rank them in terms of their likelihood of corresponding to a true DIAR.
- We design a series of visual displays and visual interactions that support intuitive exploration and validation of the generated DIAR signals.
- We evaluate the utility of MeDIAR using adverse event reports extracted from FAERS [2] with safety evaluators.

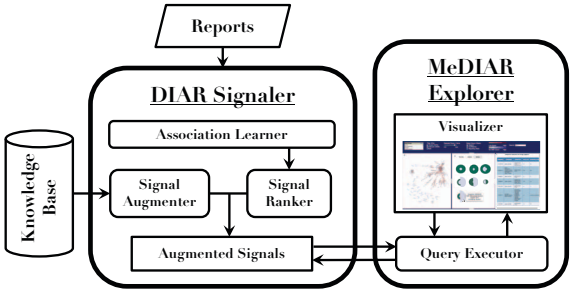

Fig. 2. The MeDIAR framework

II. The MeDIAR Framework

The MeDIAR framework in Fig. 2 consists of a back-end **DIAR Signaler** and a front-end **MeDIAR Explorer** module. To generate DIAR signals, the reports are prepared and fed into the *Association Learner* that produces non-spurious associations modeling possible DIARs in the database. The *Signal Ranker* computes their *contrast* scores and thereafter ranks them accordingly. These ranked signals are processed by *Signal Augmenter* supplementing valuable meta information extracted from the medical repository for each signal. These *Augmented Signals* form the core DIAR findings derived from the ADR reports. These findings can be accessed through a multi-layer interactive *Visualizer*. The visual interactions are translated into queries executed by *Query Executor*.

III. Key Innovations of MeDIAR

The MeDIAR system encompasses several innovations that form the foundation for the effective DIAR signal detection, exploration and validation. Our key technical contributions, namely non-spurious DIAR signal generation strategy [7], *contrast* score for ranking DIAR signal [7], and interactive MeDIAR visual analytics are introduced below.

A. Generating Non-spurious DIAR Signal.

Traditional ARL techniques model every possible combination of items that appears in a transaction. This results in a huge amount of *redundant* [8] associations. Some associations can be misleading in the context of signaling DIAR from ADR reports [7]. Consider an ADR report $t_i \equiv \mathcal{D}_i \cup \mathcal{A}_i$ with a set of suspected drugs $\mathcal{D}_i \equiv \{d_1,d_2,d_3\}$ and a set of reported ADRs $\mathcal{A}_i \equiv \{a_1,a_2\}$. This particular ADR report establishes the association between $\mathcal{D}_i$ and $\mathcal{A}_i$, expressed by the association $\mathcal{R}_1 \equiv (d_1 \wedge d_2 \wedge d_3) \Rightarrow (a_1 \wedge a_2)$. However, based upon this single report, traditional ARL would exhaustively generate 24 variants of Drug-ADR associations $((3^2-1)\times(2^2-1))$, such as $(d_1 \wedge d_2) \Rightarrow (a_1)$, $(d_1 \wedge d_3) \Rightarrow (a_2)$

etc. including $\mathcal{R}_1$. All of these associations, except $\mathcal{R}_1$, are **partial interpretations** of the report, randomly leaving out certain item(s), e.g., some drugs or some ADRs mentioned in the report. In many scenarios, these associations could be misleading unless there is additional evidence to support them. For example, $\mathcal{R}_2 \equiv d_1 \Rightarrow a_2$ tells us that taking d_1 might lead to a_2. This may however not be true in our context since this report does not *explicitly indicate* that drug d_1 by itself will lead to ADR a_2. Therefore, it cannot be confirmed based on this one ADR report.

To prune such cases when generating DIAR signals using ARL, we first define two types of valid Drug-ADR association sufficient for DIAR modeling [7], namely **explicitly supported** and **implicitly supported** Drug-ADR associations.

Definition 1: A Drug-ADR association $\mathcal{R} \equiv \mathcal{D} \Rightarrow \mathcal{A}$ is **explicitly supported** by a collection of ADR reports $\mathcal{T}$ if there exists at least one report $t_i \in \mathcal{T}$ where $t_i \equiv \mathcal{D}_i \cup \mathcal{A}_i$ such that $t_i \equiv \mathcal{D} \cup \mathcal{A}$.

If a Drug-ADR association is *explicitly* supported, according to definition 1, at least one report must exist that refers exactly to drugs and ADRs in the association and no additional ones.

Definition 2: A Drug-ADR association $\mathcal{R} \equiv \mathcal{D} \Rightarrow \mathcal{A}$ is **implicitly supported** by a collection of ADR reports $\mathcal{T}$ if there exist at least two ADR reports $t_i, t_j \in \mathcal{T}$ where $i \neq j, t_i \not\equiv t_j, t_i \equiv \mathcal{D}_i \cup \mathcal{A}_i$ and $t_j \equiv \mathcal{D}_j \cup \mathcal{A}_j$ such that $t_i, t_j \not\equiv \mathcal{D} \cup \mathcal{A}$, $\mathcal{D} \equiv \mathcal{D}_i \cap \mathcal{D}_j$ and $\mathcal{A} \equiv \mathcal{A}_i \cap \mathcal{A}_j$.

According to Definition 2, if a Drug-ADR association is *implicitly* supported, it models an association between a commonly prescribed drug combination and caused ADRs suggested by at least two reports yet it is not *explicitly* supported.

To learn these non-spurious DIAR signals from the surveillance database instead of performing post filtering, we propose an efficient mining algorithm. This algorithm maps the problem of mining valid types of signals to mining *closed* Drug-ADR associations [7].

TABLE I
EXAMPLE OF A CONTEXTUAL ASSOCIATION CLUSTER

$\mathcal{R}$	[Furosemide] [Isosorbide] [Aspirin] ⇒ [Myocardial Infarction]
$\tilde{\mathcal{R}}^2$	$\tilde{\mathcal{R}}_1^2 \equiv$ [Furosemide] [Isosorbide] ⇒ [Myocardial Infarction]
	$\tilde{\mathcal{R}}_2^2 \equiv$ [Furosemide] [Aspirin] ⇒ [Myocardial Infarction]
	$\tilde{\mathcal{R}}_3^2 \equiv$ [Isosorbide] [Aspirin] ⇒ [Myocardial Infarction]
$\tilde{\mathcal{R}}^1$	$\tilde{\mathcal{R}}_1^1 \equiv$ [Furosemide] ⇒ [Myocardial Infarction]
	$\tilde{\mathcal{R}}_2^1 \equiv$ [Isosorbide] ⇒ [Myocardial Infarction]
	$\tilde{\mathcal{R}}_3^1 \equiv$ [Aspirin] ⇒ [Myocardial Infarction]

B. Scoring and Ranking DIAR Signals.

To measure if a Drug-ADR association encodes a strong signal that indicates a severe DIAR, two factors need to be taken into consideration. First, how strong the association of ADRs is with the drug combination and second how strong the association of ADRs is with the individual or subset of drugs. If ADRs are caused by the interaction of a drug combination then not only must the ADRs be strongly associated with the drug combination but also any subset of these drugs should only be weakly associated with the particular ADRs.

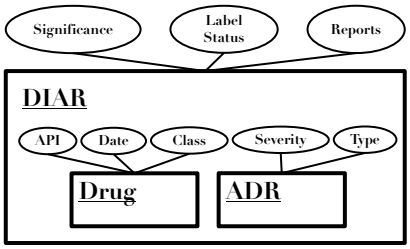

Fig. 3. The augmented DIAR signal

For the first factor, MeDIAR adopts the *confidence* model that represents a maximum likelihood estimate of the conditional probability $P(\mathcal{A}|\mathcal{D})$ for a Drug-ADR association $\mathcal{R}$. It models the strength of the association between the antecedent and consequent. High *confidence* indicates strong association while low *confidence* indicates weak association. For the second factor, we evaluate a Drug-ADR association $\mathcal{R}$ with its *contextual association cluster* (CAC) [7]. A CAC (Table I) includes a target association $\mathcal{R}$ that represents the DIAR signal along with all its contextual associations that represent the associations between the target ADRs and the subsets of the target drugs. The DIAR signal is strongest if the target association has high *confidence* and all of its contextual associations in the cluster have low *confidence*. To quantify such a *contrast* that captures the intuition of the DIAR phenomenon, we propose the *contrast* [7] measure which reflects the above observation.

C. Visual Analytics For Exploring Augmented DIAR Signals.

To provide information needed for the drug review life cycle, the generated DIAR signals are augmented with meta information (depicted in Fig 3) extracted from medical repositories. The MeDIAR visual analytics framework features an augmented DIAR signal exploration at both the *macroscopic* and *microscopic* level. At the *macroscopic* level, the interactions between all drugs are revealed via a *network* [9] visualization. We introduce a middle layer, a *glyph* representation, which enables drug review evaluators to prioritize a drug or its DIAR signal from a set of drugs that are assigned to them to reflect the drug review process. At the *microscopic* level, each candidate interaction can be viewed in context of the raw reports associated with it for further validation. When an evaluator has vetted a particular DIAR, they can report it as an actionable finding or suggest to continue monitoring it using future reports. MeDIAR's visual analytics approach to the drug review screening fills a tangible need in the early detection of severe DIARs. It allows the drug evaluator in decision making whether a DIAR signal warrants action by interactive exploration and prioritization of a signal based on its interestingness and severity, and validation by analyzing the underlying reports.

IV. DEMONSTRATION PLAN

In our demonstration, the audience will be able to interact with MeDIAR via an intuitive web interface to understand how

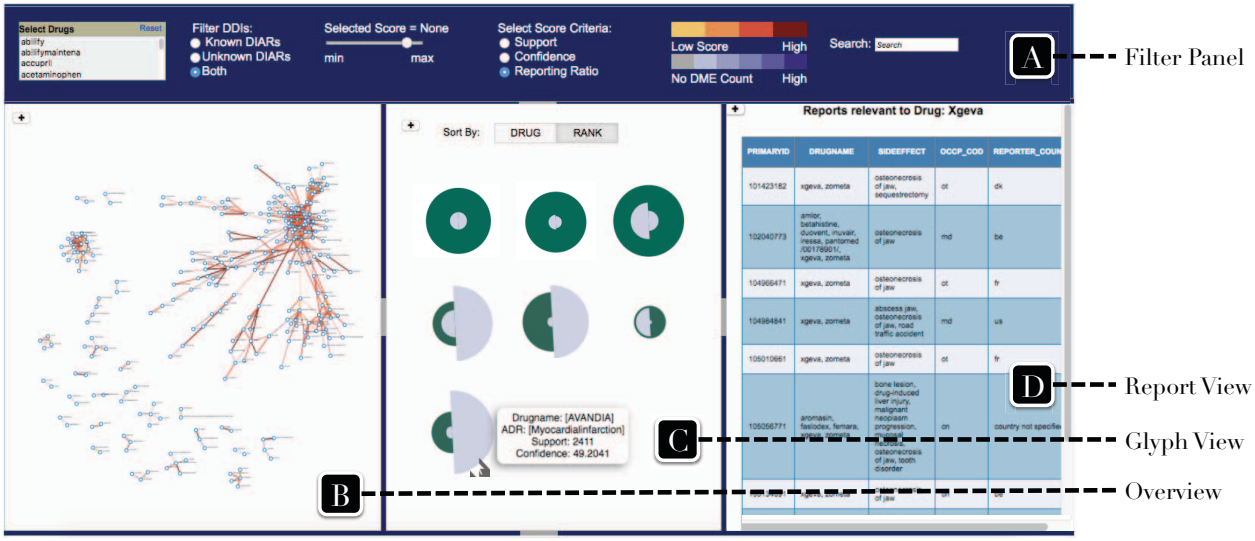

Fig. 4. The MeDIAR visual analytics

it assists drug safety evaluators in finding unknown DIARs, i.e., DIARs that have not been discovered yet, from millions of ADR reports. We conduct association mining on the adverse drug reaction reports collected from the FAERS database [2]. Our interface empowers evaluators to easily discover unknown DIARs from these reports, as described below.

Signal Generation. MeDIAR fetches the selected ADR reports directly from FAERS and prepares the raw reports into a format for the signal generation algorithm. MeDIAR mines non-spurious DIAR signals modeling the interactions among drugs. The evaluator can filter the generated signals with various criterias available on **Panel A** (Fig. 4). The selected signals are presented in a *macroscopic network* view on **Panel B** (Fig. 4) where drugs are represented by the vertices and the interactions are modeled by the edges. The width of the edge encodes the strength of the interaction quantified by the *contrast* score.

Signal Scoring Via Context Grouping. While the signals are being generated, their *significance* (*contrast* score) are computed. A CAC which is used to score its corresponding DIAR signal is visualized by *glyph* presentation in **Panel C** (Fig. 4). The green circle of a *glyph* shows the *confidence* of the target association in the CAC, i.e., the association of combination of drugs with a set of ADRs. The number of the gray sectors depicts number of contextual associations in a CAC. The radius is mapped to the *confidence* of each contextual association. In our *glyph*, the larger the green circle and the smaller the gray sectors are, the higher the score of the group is. Such deviation is measured by the *contrast* score.

Signal Augmentation and Ranking. The generated signals can then be ranked based on their *contrast* score. Signals can be sorted based on different criteria, including their rank, via **Panel C**. They are further augmented with meta information, e.g. *drug class*, *severity*, *label status*, *links to the original reports* etc., extracted from the medical repositories. This meta data is used to improve the signal's interpretation. It provides additional information for further investigation. In particular, the unknown versus previously known status of a DIAR can be filtered via **Panel A**, so that the evaluators can quickly narrow their attention onto the unknown severe DIAR signals. In **Panel C**, all DIAR signals related to a selected drug from **Panel B** can be displayed.

DIAR Signal Confirmation Via Reports. By hovering over a DIAR represented by a *glyph* in **Panel C**, the mouse-over shown in white displays the details of the particular contextual association. A click then allows us to select all relevant ADR reports, which would be retrieved from the surveillance database and presented in a report listing panel (**Panel D** Fig. 4) providing a *microscopic* level of signal investigation. The evaluator is now able to read, highlight and make notes on the reports for further signal validation.

REFERENCES

[1] T. Chan, "Adverse interactions between warfarin and nonsteroidal anti-inflammatory drugs: mechanisms, clinical significance, and avoidance." *The Annals of pharmacotherapy*, vol. 29, no. 12, pp. 1274–1283, 1995.

[2] FDA Adverse Event Reporting System (FAERS) data statistics. https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm. Accessed: 2017-09-29.

[3] Canada Vigilance Adverse Reaction Online Database. http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-eng.php. Accessed: 2017-06-09.

[4] L. Wei and J. Scott, "Association rule mining in the us vaccine adverse event reporting system (vaers)," *Pharmacoepidemiology and Drug Safety*, vol. 24, no. 9, pp. 922–933, 2015.

[5] R. Harpaz, H. S. Chase, and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems," *BMC Bioinformatics*, vol. 11, no. S-9, p. S7, 2010.

[6] R. J. Bayardo Jr and R. Agrawal, "Mining the most interesting rules," in *Proceedings of the 5th ACM SIGKDD*, 1999, pp. 145–154.

[7] X. Qin, T. Kakar, S. Wunnava, E. A. Rundensteiner, and L. Cao, "Maras: Signaling multi-drug adverse reactions," in *Proceedings of the 23rd ACM SIGKDD*, 2017, pp. 1615–1623.

[8] M. J. Zaki, "Generating non-redundant association rules," in *Proceedings of the 6th ACM SIGKDD*, 2000, pp. 34–43.

[9] R. Ball and T. Botsis, "Can network analysis improve pattern recognition among adverse events following immunization reported to vaers?" *Clinical Pharmacology & Therapeutics*, vol. 90, no. 2, pp. 271–278, 2011.