Causality driven data integration for adverse drug reaction discovery

Chen Wang, Sarvnaz Karmi

CSIRO Computational Informatics

SUMMARY

We describe an ongoing effort in CSIRO for partially automating causality discovery in the Adverse Drug Reaction (ADR) detection process. The proposed method integrates data from multiple sources based on rules that indicate causality.

INTRODUCTION

Drug adverse reactions are a major threat to public health and impose huge costs to healthcare systems. Postmarketing surveillance aims to reduce the effects of adverse drug events. There are two types of ADR discovery systems operating around the globe: passive discovery and active discovery. They differ in the data used for detecting unexpected harm caused by the normal use of a drug at the normal dosage as per label or prescription. Passive ADR discovery, which has been established for decades, uses reports that are voluntarily submitted by pharmaceutical companies, healthcare professionals and consumers. Regulatory bodies, such as TGA (Therapeutic Goods Administration) and FDA (Food and Drug Administration), maintain very large databases of such reports which they use to mine the potential safety signals. More recently, active discovery has been introduced. Active discovery monitors healthcare data from a variety of sources such as electronic health records, health insurance claims, medical literature, or even recently medical forums to identify potential signals automatically using text and data mining techniques. Active discovery intends to discover unexpected adverse events as early as possible and is therefore also called near real-time drug safety surveillance. An example of such system is recently proposed in FDA Sentinel initiative which relies on sharing deidentified patient data among a number of organisations.

Both types of ADR discoveries ultimately lead to establishing the causal relationship between a drug and unexpected adverse reactions. Often ADR discovery starts with data-mining techniques for disproportionality detection of the reports about a drug and an adverse reaction in comparison to other pairs of drugs and adverse reactions. These potential ADRs are then examined in medication safety review and assessment meetings. The main task of these meetings is to establish the causality between a drug and its adverse reactions. This is largely a manual process in the current practice and often generates wide variability in assessment[1,2,3]. Even though the shortcomings of the current process were recognised in 70s[1], there is not much improvement in the practice of establishing causality between a drug and an ADR so far. As ADR related data become increasingly accessible in electronic format and with the increase in processing power and techniques of dealing with big data, it is now possible to introduce carefully designed algorithms to assist the causality reasoning process and therefore automate some of the manual steps in this assessment to reduce variability. We note that the current process, endorsed by WHO, is still largely based on Naranjo's questionnaire[1] designed in 1981. To achieve this, there are two major requirements: first, it is essential to understand and capture the reasoning process in the existing practice. A good reasoning process tends to minimise the variability and inconsistency in assessment as shown in[1]. Second, integrating data from various sources is essential for reaching correct conclusions in the reasoning process, e.g. additional data about background of the patients in ADR reports may help to identify causes of an ADR. This is of course only possible with collaboration of multiple health agencies to make such data accessible. Below, we propose a causality detection method to address these requirements.

DESCRIPTION

Previous work trying to establish causality between a drug and its unexpected adverse reactions used a well designed questionnaire to guide the assessment process[1]. The answers of these questions were assigned different scores and the total score of each rater determines the certainty of the rater on whether a drug D causes a reaction R. A consensus among raters served as an indicator of the causality of D and R.

Our proposed method contains two steps: (1) Design rules to capture the causality reasoning process using domain-expertise and the current known knowledge of ADRs per each drug or active ingredient; and (2) Process different data sources based on these rules to establish if a given drug D causes a specified adverse reaction R.

A starting point for rule identification is using the existing questionnaires, and also formalising the reasoning process within the review and assessment teams inside the regulatory. For instance, consider a specific drug D and its possible adverse reaction R and a given dataset S (e.g, electronic health records and clinical notes). The following rules could be considered for causality discovery:

Dr Chen Wang

Senior Research Scientist
CSIRO Computational Informatics

chen.wang@csiro.au

Dr Chen Wang is a Senior Research Scientist at CSIRO Computational Informatics. He received his PhD from Nanjing University. His research interests are primarily in distributed, parallel and trustworthy systems. His current work focus on data analytics systems for drug adverse reaction discovery. His recent work include accountable distributed systems and cloud computing. He is also an Honorary Associate of the School of Information Technologies at the University of Sydney. Dr Chen Wang has industrial experience. He developed a high-throughput event delivery system and a medical image archive system, which are used by many hospitals and medical centres in USA.

1. Discontinued D, R still existed;
2. Discontinued D, R improved;
3. Readministered D, R reappeared;
4. Increased the dose of D, R became more severe;
5. Decreased the dose of D, R became less severe;
6. Factors F1, F2 and F3 cause R.

These rules capture common reasoning used in identifying whether D causes R. The set of rules are extensible. With these rules defined, the next step is to process data based on these rules to discover causality between a drug and a given adverse reaction. In order to achieve this, we first build a data model that contains necessary data fields required by these rules. For example, to support the rules above, we need information about actions taken on a drug by a consumer, or instructions of a medical professional to the patient, such as the discontinuing its use, changing its dose etc. as well as additional information about other factors that may cause the adverse reaction. After the rule list is completed, a table is constructed to capture the data model. See Table 1 for an example. The headers of Column 2 to Column 6 show a sample data model. Afterwards, we process each data source using information extraction techniques and assisted by medical ontologies and drug knowledge repositories to populate the table. The last column "D causes R" in Table 1 represents the decisions and is partially populated via existing knowledge.

Table 1. Summary of reports regarding drug D and suspected adverse reaction R

REPORT	D DISCONTINUED	D READMINISTRATERED	DOSE CHANGE	OTHER FACTORS	R CHANGE	D CAUSES R
1	Yes	N/A	No	None	Improved	Yes
2	Yes	N/A	No	Unknown	No improvement	Unknown
3	Yes	N/A	No	F2	No improvement	No
4	No	No	Decreased	F3	Improved	Yes
5	No	No	Decreased	Unknown	No improvement	Unknown

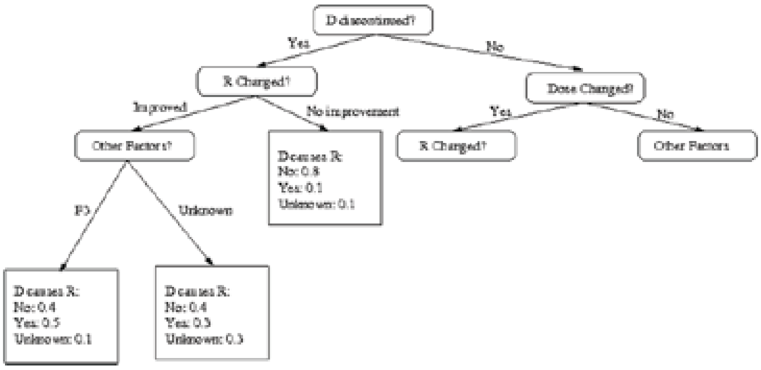

Figure 1. A sample decision tree for ADR causality discovery

Based on the table, we use a decision tree to classify the data. Human annotated data are used as the training set. A sample decision tree classifier is shown in Figure 1. Note that this tree only partially covers Table 1. The final decisions on causality (Yes, No, or Unknown) will be based on a threshold on the probabilities generated by decision. The decision tree evolves as the number of confirmed causality pairs increases. As the data model is independent of underlying data sources, our method is capable of dealing with multiple data sources as long as they contain at least some of information needed by the data model.

CONCLUSION

Causality discovery is essential to detect potential adverse drug reactions. However, the implementation challenges are extracting high quality causality information from a variety of data and dealing with different level of credibility of information from different data sources.

REFERENCES

1. C. Naranjo, U. Busto, E. Sellers, P. Sandor, I. Ruiz, E. Roberts, E. Janecek, C. Domecq, and D. Greenblatt. A method for estimating the probability of adverse drug reactions. Clinical Pharmacology and Therapeutics, 30:239–245, 1981.
2. R. P. Naidu. Causality assessment: A brief insight into practices in pharmaceutical industry. Perspect Clin Res 2013;4:233-6.
3. N. Anderson, J. Borlak. Correlation versus causation? Pharmacovigilance of the analgesic flupirtine exemplifies the need for refined spontaneous ADR reporting. PLoS One. 2011;6(10):e25221