NSIDES: DRUG EFFECT DISCOVERY USING THE FDA ADVERSE REPORTING SYSTEM

RAMI S VANGURI

Department of Biomedical Informatics, Columbia University,
New York, NY 10032 USA
E-mail: r.vanguri@columbia.edu

JOSEPH D ROMANO

Department of Biomedical Informatics, Columbia University,
New York, NY 10032 USA
E-mail: jdr2160@columbia.edu

TAL LORBERBAUM

Department of Biomedical Informatics, Columbia University,
New York, NY 10032 USA
E-mail: tal.lorberbaum@columbia.edu

VICTOR NWANKWO

Department of Biomedical Informatics, Columbia University,
New York, NY 10032 USA
E-mail: vtn2106@columbia.edu

CHOONHAN YOUN

San Diego Supercomputer Center, University of California, San Diego,
La Jolla, CA 92093 USA
E-mail: cyoun@sdsc.edu

NICHOLAS P TATONETTI

Department of Biomedical Informatics, Columbia University,
New York, NY 10032 USA
E-mail: nick.tatonetti@columbia.edu

Adverse drug events are a leading cause of morbidity and mortality around the world. Regulatory agencies, such as the Food and Drug Administration (FDA), maintain large collections of adverse event reports, providing an opportunity to retrospectively study drug and drug combination effects. We mined the FDA Adverse Event Reporting System (FAERS) for significant adverse reactions and developed a database of drug effects, known as nSides. FAERS contains millions of reports covering thousands of drugs and thousands of effects, requiring the computing of approximately X billion models. In order to calculate side effect significances for all drug combinations reported in FAERS, we present a distributed, on-demand computational architecture.

1. Introduction

Spontaneous reporting systems such as the FDA Adverse Event Reporting System (FAERS) are important resources for detecting drug adverse events after a drug is approved (pharmacovigilance). However, pharmacovigilance algorithms often lead to many false positive and false negative findings due to issues of confounding, and detection of drug-drug interactions is an even greater challenge. We previously developed databases for off-label drug effects (Offsides) and drug interactions (Twosides) that account for these limitations using a novel Statistical Correction for Uncharacterized Bias (SCRUB).[1] We re-mined FAERS with an updated algorithm to populate a new version of the databases, known as nSides. nSides also contains a public web gateway (http://nsides.io/) accessible to researchers, clinicians and patients alike.

2. Data Sources

There are several data sources which are involved in nSides. We use a curated version of the FDA Adverse Event Reporting System (FAERS) known as Adverse Event Open Learning through Universal Standardization (AEOLUS).[2] AEOLUS aims to clean and normalize the data by removing duplicate cases. This is done by applying standardized vocabularies in the form of RxNorm to map drug names and SNOMED-CT to map outcomes. The AEOLUS dataset is publicly available.

3. Methods

3.1. *Algorithm*

The algorithm used to develop the databases used for nSides is an updated version of the one used to populate the Offsides and Twosides databases. These databases contain side effect significances calculated using raw FAERS data.[1] Generally, a standard signal detection algorithm involves conducting a disproportionality analysis by comparing the observed reporting frequency of a drug and outcome to the expected reporting frequency of all other drugs and the outcome. The metric is known as a Proportional Reporting Ratio (PRR). If the outcome occurred by chance, the frequencies will be equal and the PRR will be one. If the PRR is much larger than one, the null hypothesis is rejected. To reduce sampling variance and selection bias, propensity score matching is implemented to form the groups used in the disproportionality analysis. This procedure, known as SCRUB, matches cases and controls between patients exposed and not exposed to a particular drug (Offsides) or two drugs (Twosides) to mitigate confounding biases.

There are several key differences between the Offsides and Twosides databases and nSides. The updated algorithm uses a deep learning model instead of logistic regression to calculate propensity scores to match cases and controls.

3.2. *Computational Challenge*

The AEOLUS dataset contains ≈4,500 drugs, ≈4,500,000 reports and ≈7,500 effects. In order to populate nSides, a unique model needs to be created for each drug present in AEOLUS, presenting a computational challenge. This becomes even more intractable when considering drug interactions.

References

1. N. P. Tatonetti, P. P. Ye, R. Daneshjou and R. B. Altman, *Science Translational Medicine* **4**, 125ra31 (2012).
2. J. M. Banda, L. Evans, R. S. Vanguri, N. P. Tatonetti, P. B. Ryan and N. H. Shah, *Scientific data* **3** (2016).