Wikipedia Drug Safety Advisory Committee: Distilling A Drug Adverse Effect Reference Set Using Wisdom of The Crowd

Yonatan Bilu and Chen Yanover
KI Research Institute

The Israeli Institute for Applied Research in Computational Health

Background

Datasets describing potential adverse drug effects (ADEs) can roughly be separated into two types: manually curated, typically, small ones, and large ones, which are automatically extracted from relevant raw data, such as package inserts (SIDER[1]) or collected adverse effects reports (OFFSIDES[2]). However, these obtained lists often mix rare and common ADEs alongside those caused by misuse or abuse of drugs and, therefore, using such datasets as a ground truth for prediction (or estimation) algorithms may be misleading. Hence, it would be useful to distil from such datasets those ADEs which are frequently encountered in practice.

One way to do this is for expert medical professionals to manually curate a large dataset and pinpoint, based on their medical experience, adverse effects commonly observed in practice. However, this can be potentially biased, and is usually costly and time consuming – especially if one wants to integrate the experiences of several medical professionals. An alternative, which encapsulates the same motivation, is to cross-reference the adverse effects listed in a dataset, with those mentioned in Wikipedia. Indeed, Wikipedia articles often reflect the aggregated knowledge of multiple editors, many of which are often well versed in the field on which they write.

Here we apply natural language processing (NLP) methods on Wikipedia drug articles to verify SIDER and OFFSIDES ADEs and come up with an improved ADE list. We demonstrate the superior accuracy of the obtained set using ground truth data from published clinical trials as well as manually-constructed ones.

Methods

Given a name of a drug, we retrieve the corresponding Wikipedia article, using Wikipedia's API, and extract the section describing the drug's adverse effects as well as the opening section.

An ADE mentioned in SIDER or OFFSIDES is considered "verified" by Wikipedia if it appears in one of the sentences extracted as above. We examine only ADEs which belong to the "preferred term" hierarchy level in MedDRA, and consider an ADE as "verified" also if one of its "lower-level terms" in the hierarchy is mentioned. We discard ADEs which are listed in SIDER as indications, or if they are mentioned in the first paragraph of the Wikipedia article's summary section, since this paragraph typically mentions indication.

We evaluate the sensitivity and specificity of the distilled adverse event set, compared to the original one, on a ground truth set derived from clinical trials, using the code of Steinberg et al.[3], and requiring that one of the treatment arms is placebo (leaving all other parameters unchanged). The clinical trial set includes 412 positive controls. Additionally, we examine how well aligned these lists are with the reference set of Voss et al.[4], which lists 208 drug-ADE pairs for a nine ADEs. For all evaluations, we report precision, recall, and F1 score.

References

1. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016
2. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012
3. Steinberg E, Yadlowsky S, Shah NH. A Clinical Trial Derived Reference Set for Evaluating Observational Study Methods.
4. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. J Biomed Inform. 2017

Results

SIDER includes 141,311 drug-ADE pairs (1,344 drugs, 4,563 ADEs); of these, 9,590 (6.8%) are Wikipedia-verified (776 drugs, 875 ADEs). From OFFSIDES we extracted 1,051,942 pairs, based on their PRR statistics (2381 drugs, 9957 ADEs); of which 9820 (0.9%) are Wikipedia-verified (1,013 drugs, 1,229 ADRs).

Ground truth	Reference Set	Precision	Recall	F1
Clinical Trials[3]	SIDER	0.11	0.88	0.19
	SIDER-Distilled	**0.87**	**0.72**	**0.76**
	OFFSIDES	0.04	0.59	0.07
	OFFSIDES-top	0.09	0.11	0.10
	OFFSIDES-Distilled	**0.57**	**0.49**	**0.53**
Voss et al.[4]	SIDER	0.002	0.21	0.004
	SIDER-Distilled	**0.009**	**0.10**	**0.017**
	OFFSIDES	0.0006	0.14	0.001
	OFFSIDES-top	0.0019	0.02	0.004
	OFFSIDES-Distilled	**0.008**	**0.08**	**0.014**

Macro-averaged Precision, Recall and F1-scores. The Distilled sets are those obtained via the method described here. Drug-ADE pairs in offside are scored therein with a proportional reporting ratio score (PRR). The OFFSIDES-top set consists, for each drug, of the top *k* scoring ADEs according to this score, with *k* being equal to the number of ADEs verified by Wikipedia for the drug.

Conclusions

We present an automatically-curated dataset of drug-ADEs, focusing at this time on positive examples. This dataset can be useful not only as a benchmark for the prediction of adverse ADEs from EHRs, but also for general prediction and inference methods. Exhibiting good performance on this dataset is indicative of the method's power, and suggests that it could work well when performing other types of predictions, for which curated data is scarce or unavailable – such as the adverse effects of vaccines.

We hope that this method can be integrated as a module into OHDSI's Common Evidence Model, as an annotation layer over the datasets therein.

Contact: {yonatan, chen}@kinstitute.org.il