Wikipedia Drug Safety Advisory Committee: Distilling a Drug Adverse Effect Reference Set Using Wisdom of the Crowd

Yonatan Bilu, Chen Yanover

Background

Comparative reference sets have been developed to support evaluation of methodologies in observational studies[1,2]. Such sets, indicating the (direction of) effect of one drug ingredient compared to another on a given outcome, have been successfully used to assess the performance of multiple comparative effectiveness methods, e.g. in Ref 3. However, some methods – importantly, estimating the efficacy of a vaccine or the safety of drugs – consider a no-treatment arm, and their evaluation requires reference sets that compare the effect of a treatment on an outcome, relative to no-treatment (or placebo). The common practice in the OHDSI community is to identify negative controls, either manually or from literature, product labels, and spontaneous reports using ATLAS[4,5]; then, derive synthetic positive controls from the obtained negative controls. Alternatively, controls can be obtained from knowledgebases, such as the SIDER dataset[6], which extracted adverse drug effects (ADEs) from public documents and package inserts using natural language processing (NLP) methods, or OFFSIDES[7] dataset, which is based on the FDA Adverse Event Reporting System. However, the obtained lists typically 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 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. We hope that our distilled list serves the community to systematically test placebo-based study designs and methodologies.

Methods

Distilled set. To identify ADEs mentioned in a drug's Wikipedia article we rely on two simplifying observations in our setting: (1) We do not need to identify all ADEs, only to determine which of those mentioned in SIDER also appear in the Wikipedia article; and (2) such articles usually mention adverse effects either in the initial summary, or in a dedicated section.

Given a name of a drug, we retrieve the corresponding Wikipedia article, and process it as follows: we search for sections whose title contains one of the phrases: "side effect", "side-effect" or "adverse". If such a section is found, it is parsed into subsections and then into sentences. Subsections are discarded if they contain any of the words - 'tolerance', 'dependence', 'withdrawal', 'interactions', 'interaction', 'overdose', 'discontinuation', since we are interested in adverse effects with respect to placebo. In addition, the opening summary section is extracted, and parsed into paragraphs and sentences.

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.

Evaluation. We first evaluated 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[2] and requiring that one of the treatment arms is placebo (leaving all other parameters unchanged). The control trial set included 412 positive controls. Additionally, we examined how well aligned these lists are with Ref 4’s reference list, which lists 208 drug-ADE pairs for a nine ADEs. For all evaluations, we report precision, recall, and F1 score.

Results

SIDER includes 141,311 drug-ADE pairs (1,344 drugs, 4,563 ADEs); of these, 9,590 are Wikipedia-verified (776 drugs, 875 ADEs). From OFFSIDES we extracted 1,051,942 pairs, based on their proportional reporting ratio (PRR) statistics (2381 drugs, 9957 ADEs); of which 9820 are Wikipedia-verified (1,013 drugs, 1,229 ADRs). The much smaller distilled lists obtain, expectedly, lower recall but its precision and F1 score are higher, in the comparisons with both ground truth sets (Table 1). To better illustrate the effectiveness of using Wikipedia for this task, we also compare the sub-list of highest-PRR pairs in OFFSIDES, of the same length as the Wikipedia-distilled one, to both ground-truths.

Ground truth	Reference Set	Precision	Recall	F1
Clinical Trials	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
Manual	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
	Wiki-Distilled	0.008	0.08	0.014

Table 1 Evaluation of SIDER, OFFSIDES and their distilled set, vs. two ground truths, derived from Clinical Trials[2] and manual curation[4]. All measures are macro-averaged over the set of drugs.

Conclusion

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.

Importantly, this dataset is oriented towards precision – it does not aim to capture all ADEs that a drug may cause. Rather, it lists a subset of these ADEs, for which confidence is high. Hence, evaluating a model vs this benchmark should be done accordingly.

In ongoing work, we indeed aim to evaluate this dataset versus predictions deduced from EHRs. Initial research in this direction, using OHDSI's SelfControlledCaseSeries package[8], and naively labeling drug-ADE pairs according to the sign of their log ratio scores, revealed two inherent problems: (1) In high-scoring pairs, the effect is almost always an indication rather than an ADE; and (2) effects common in the data tend to get a high score, even when we could find no external corroboration for them being an ADE. Addressing these problems will also allow to evaluate a set of negative controls which we have constructed using similar methods.

Finally, we believe that the method developed here may have broader use. We showed that a rather naïve use of Wikipedia to curate medical data can greatly improve the accuracy of the results. It is likely that this can be extended to other types of data.

References

1. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a Reference Set to Support Methodological Research in Drug Safety. *Drug Saf*. 2013;36(S1):33-47. doi:10.1007/s40264-013-0097-8

2. Steinberg E, Yadlowsky S, Shah NH. A Clinical Trial Derived Reference Set for Evaluating Observational Study Methods. *ArXiv200614102 Stat*. Published online June 24, 2020. Accessed July 20, 2020. http://arxiv.org/abs/2006.14102

3. Ryan PB, Stang PE, Overhage JM, et al. A Comparison of the Empirical Performance of Methods for a Risk Identification System. *Drug Saf*. 2013;36(S1):143-158. doi:10.1007/s40264-013-0108-9

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;66:72-81. doi:10.1016/j.jbi.2016.12.005

5. OHDSI. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI; 2019. https://books.google.co.il/books?id=JxpnzQEACAAJ

6. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44(D1):D1075-D1079. doi:10.1093/nar/gkv1075

7. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012;4(125):125ra31. doi:10.1126/scitranslmed.3003377

8. Schuemie M, Ryan P, Shaddox T, Suchard M. *SelfControlledCaseSeries: Self-Controlled Case Series*.; 2021. https://github.com/OHDSI/SelfControlledCaseSeries