Drug Saf (2015) 38:113–114
DOI 10.1007/s40264-014-0245-9

LETTER TO THE EDITOR

Comment on: "Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance"

Rave Harpaz · William DuMouchel · Nigam H. Shah

Published online: 29 November 2014

To the Editor:

We read the article by Norén et al. [1] with great interest and commend their effort in bringing forward a critical issue in the evaluation of signal detection methodologies, namely the choice of a benchmark (a reference standard) and an associated evaluation strategy.

Norén et al. argue that signal detection is fundamentally a prognostic activity. Therefore, evaluation strategies should aim to emulate a prospective analysis of signal detection in lieu of a commonly applied yet unsatisfactory approach of retrospective analysis based on well established associations such as those comprising the Observational Medical Outcome Partnership (OMOP) [2] and EU-ADR benchmarks [3]. Norén et al. demonstrate that the two evaluation strategies may lead to different conclusions. They partially attribute this discrepancy to biasing effects (e.g., the influence of publicity on spontaneous reporting and on patient management), which are a consequence of examining well established associations in a retrospective manner. Taken together, Norén et al. argue that evaluations should be based on benchmarks consisting of emerging or recently labeled adverse drug reactions (ADRs), which are to be applied in a manner that simulates prospective analysis by backdating the analyses to periods prior to the conception of these ADRs.

We agree with the issues raised by Norén et al., but do not go as far as the dismissal of existing benchmarks. In an effort to shed light over this debate we recently created a time-indexed benchmark specifically designed to support the type of prospective evaluations proposed by Norén et al. The benchmark consists of recently labeled adverse events communicated by the US FDA in 2013. It includes 62 positive controls and 75 negative controls, covering 38 adverse events and 44 drug ingredients. Together with its description, the benchmark is available through Nature Scientific Data [4]. A preliminary investigation that applied this benchmark to evaluate FDA Adverse Event Reporting System (FAERS)-based signal detection provides support for the argument by Norén et al., in contrast with our earlier study based on the OMOP benchmark [5].

Despite these results we maintain our view that the two approaches should supplement each other. A key advantage to using well established positive controls is in the reliability of their supporting evidence. In a benchmark created prior to the inception of a given recently labeled or emerging ADR, it is possible that this 'true' ADR would have been classified as a negative control. Likewise, the status of a recently labeled ADR (positive control in some benchmarks) may be revised based on new refuting evidence. Thus, the increased level of uncertainty associated with experiments based on such recently labeled or emerging ADRs cannot be ignored.

Another issue is that many post-approval adverse events emerge shortly after a drug is introduced to the market. This short duration suggests that a backdated prospective analysis of benchmarks containing newly introduced drugs (an important target for monitoring) may not be feasible given that an insufficient amount of data will be available for analysis. In such cases, a retrospective analysis is likely the only option.

Perhaps the most important issue is the interpretation of backdated analyses. A key question that follows a backdated analysis is whether or not the conclusion drawn from

R. Harpaz (✉) · W. DuMouchel
Oracle Health Sciences, Bedford, MA, USA
e-mail: rave.harpaz@oracle.com

N. H. Shah
Stanford University, Stanford, CA, USA

the analysis can be extrapolated to present times; that is, the time in which we will actually use signal detection to monitor for new issues. Taking the example provided by Norén et al., can we safely say that their experiment backdated to the end of 2004 reflects the state of signal detection in the year 2014? Arguably not. Due to changes in policy, data collection, or coding practices, it is unlikely that the intrinsic properties of the data on which signal detection is applied remain constant over time. Unless such an experiment is repeatedly replicated in future time points, and the results of the experiment remain consistent, we cannot argue for their generalizability with confidence. The need for such repeated evaluations points to another core issue, which is that the relevance of such benchmarks is time-sensitive in itself. New sets of benchmarks containing newer ADRs will need to be continuously tracked and curated in order to use them for backdated prospective analyses.

In summary, we strongly agree with the need for additional benchmarks and support the ideas brought forth by Norén et al. Given our experience in creating and using such a time-indexed benchmark of recent ADRs, we point to the challenges associated with implementing and interpreting such benchmarks. We stress that keeping such proactive benchmarks up-to-date with new safety information requires a significant, ongoing commitment, and needs to be a community effort such as that under the Observational Health Data Science Initiative (http://www.ohdsi.org) [6].

Last but not least, the ultimate objective of signal detection is to identify new safety issues with high fidelity and in a timely manner. This suggests that the evaluation of signal detection methodologies should consist of at least one more dimension—that of time-to-detection [7]. To our knowledge, time-to-detection has yet to be accepted as an additional evaluation aspect. Here we envision an evaluation strategy that measures how early different methodologies detect signals while factoring in their false alert rates. It is conceivable that the discriminatory power of signal detection methodologies (as measured by prospective or retrospective strategies) and the time-to-detection are not positively correlated. We therefore propose that this aspect of signal detection, along with possibly cost, severity, and other triage approaches, should also be investigated as part of an overall model to evaluate the effectiveness of signal detection methodologies.

Acknowledgments Rave Harpaz and Nigam H. Shah acknowledge support by NIH Grant U54-HG004028 for the National Center for Biomedical Ontology and by NIGMS Grant GM101430-01A1.

Competing financial interests Rave Harpaz, William DuMouchel and Nigam H. Shah declare no competing financial interests. Rave Harpaz and William DuMouchel are employed by Oracle Health Sciences. Nigam H. Shah is a Science Advisor to ApixioInc (http://www.apixio.com), and Kyron Inc (http://www.kyron.com).

References

1. Noren GN, Caster O, Juhlin K, Lindquist M. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Saf. 2014;37(9):655–9.
2. 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(Suppl. 1):S33–47.
3. Coloma PM, Avillach P, Salvo F, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf. 2013;36(1):13–23.
4. Harpaz R, Odgers D, Gaskin G, et al. A time-indexed reference standard of adverse drug reactions. Nat Sci Data 1. 2014. Art ID 140043. doi:10.1038/sdata.2014.43.
5. Harpaz R, Dumouchel W, Lependu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clin Pharmacol Ther. 2013;93(6):539–46.
6. Boyce RD, Ryan PB, Noren GN, et al. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Saf. 2014;37(8):557–67.
7. LePendu P, Iyer SV, Bauer-Mehren A, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther. 2013;93(6):547–55.