Drug Saf (2014) 37:655–659
DOI 10.1007/s40264-014-0198-z

CURRENT OPINION

Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance

G. Niklas Norén · Ola Caster · Kristina Juhlin · Marie Lindquist

Published online: 9 July 2014

Abstract Pharmacovigilance seeks to detect and describe adverse drug reactions early. Ideally, we would like to see objective evidence that a chosen signal detection approach can be expected to be effective. The development and evaluation of evidence-based methods require benchmarks for signal detection performance, and recent years have seen unprecedented efforts to build such reference sets. Here, we argue that evaluation should be made against emerging and not established adverse drug reactions, and we present real-world examples that illustrate the relevance of this to pharmacovigilance methods development for both individual case reports and longitudinal health records. The establishment of broader reference sets of emerging safety signals must be made a top priority to achieve more effective pharmacovigilance methods development and evaluation.

1 Introduction

Once trapped and fenced in at the zoo, zebras are easy to spot, but in the high grass and vast expanse of the savannah they will often escape even the trained eye. Zebra-spotting performance in the zoo cannot be extrapolated to the savannah, nor can the optimal skill set, although there may be some correlation. Therefore, aspiring trappers had better abandon the zoo and seek more relevant terrain for training.

A similar situation prevails in pharmacovigilance, where our fundamental aim is to detect and describe adverse drug reactions *early*, and where there are numerous possibilities for how to do so. There are individual case reports [1], longitudinal health records [2], internet search patterns [3] and social media [4]. There is disproportionality analysis [1], regression [5, 6], adjustment by propensity scores [7, 8], self-controlled designs [2, 9] and more. Expert judgment is important in choosing methods and datasets for pharmacovigilance, but ideally we would like to see objective evidence that a chosen approach can be expected to be effective. To this end, we need benchmarks for performance evaluation. This is well-understood and broadly accepted: recent years have seen unprecedented efforts to build broad reference sets of established adverse drug reactions and adverse events without evidence for (or with evidence against) causal associations with a drug [10, 11]. If these reference sets indirectly (or directly, as in the Observational Medical Outcomes Partnership studies) drive our choice of analytical approach, then their choice of positive and negative controls is essential. In particular, we must be careful in distinguishing between emerging safety signals and established causal associations, as they are different in nature. Below, we present real-world examples where evaluation of signal detection methods against established safety signals yield fundamentally different conclusions than evaluation against emerging safety signals. We show the relevance of these considerations in pharmacovigilance methods development for both individual case reports and longitudinal health records.

G. N. Norén (✉) · O. Caster · K. Juhlin · M. Lindquist
Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Box 1051, 751 40 Uppsala, Sweden
e-mail: niklas.noren@who-umc.org

G. N. Norén
Department of Mathematics, Stockholm University, Stockholm, Sweden

O. Caster
Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden

2 Examples

As a first example, consider narcolepsy in children and adolescents after Pandemrix vaccination. This safety signal emerged in the wake of broad vaccination initiatives under the pandemic threat of 2009. As of June 2014, there are 752 reports from 12 countries on the MedDRA® [1] Preferred Term (PT) narcolepsy with Pandemrix vaccination in the WHO global individual case safety reports database VigiBase®. However, if we backdate our analysis to 17 August 2010, when the signal was first communicated to the general public [12], there were only three reports of narcolepsy after Pandemrix vaccination in VigiBase®, all originating in Sweden. In other words, while early detection of this signal in VigiBase® would require a reasonably sensitive signal detection method, use of current data might lead us to treat this as a true positive for almost any approach.

Now consider the challenge of evaluating signal detection performance against broad references of such positive and negative controls. Contemporary research has reported improved performance of multivariate analytics compared to disproportionality screening, for the analysis of individual case reports [6, 8]. This seems plausible, since the new methods offer innovations such as adjustment for comedications and indications for treatment. On the other hand, these studies have used established adverse drug reactions as positive controls in their evaluation, and for such benchmarks, simple report counts too can outperform disproportionality analysis: Fig. 1 shows the sensitivity and specificity for identifying established adverse drug reactions at different thresholds for a disproportionality measure (lower limit of a 95 % credibility interval for the Information Component (IC_{025}) [13]) and for the raw numbers of reports, respectively. Here, individual MedDRA® PTs corresponding to adverse reactions listed in section 4.8 of the summary of product characteristics (SmPC) for European centrally authorised products[2] are used as positive controls. These results show that report counts are significantly better predictors than disproportionality measures for events listed on the SmPC, and based on that one might be tempted to conclude that, as a community, we have wasted 15 years pursuing disproportionality analysis, when we would have been better off continuing to screen based on raw numbers of reports. However, this conclusion would only be valid to the extent

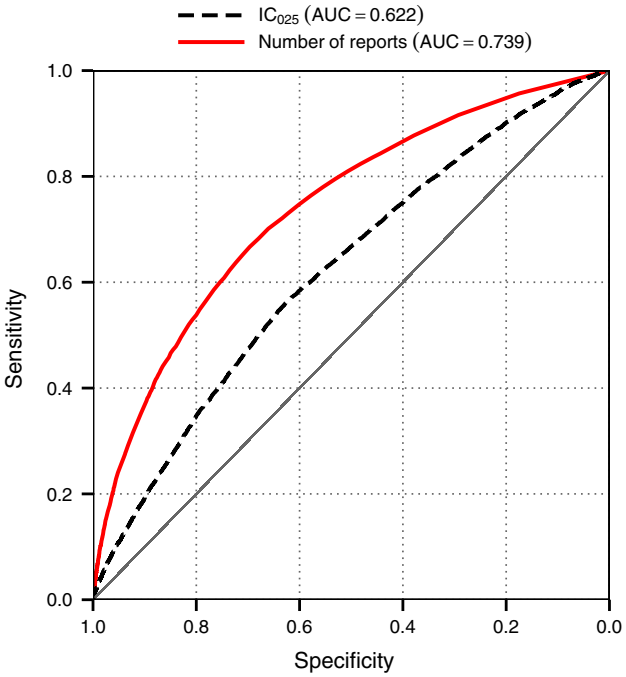

Fig. 1 Sensitivity and specificity for established adverse drug reactions of disproportionality analysis (IC_{025}) and raw report counts, respectively. The 16,811 positive controls of this reference set are individual MedDRA® PTs corresponding to well-established adverse drug reactions listed on the SmPCs for centrally authorised products in Europe; the 16,811 negative controls are drugs paired with PTs for which no other PT in the same MedDRA® High-Level Term were listed on the drug's European SmPC in 2012. Data for both positive and negative controls were from VigiBase® as of May 2013. Thresholds for the report counts yield better specificity than thresholds for the disproportionality measure with the same sensitivity. The AUC values are 0.622 for IC_{025} and 0.739 for the raw report counts ($p \ll 0.05$ according to DeLong's test). *AUC* area under the receiver operating characteristic curve, IC_{025} lower limit of a 95 % credibility interval for the Information Component, *PT* Preferred Term, *SmPC* summary of product characteristics

that the reference were fit-for-purpose, and there is evidence to suggest that it is not: Fig. 2 shows the corresponding graph for historical safety signals from the European Medicines Agency (EMA) [14] backdated to the time around the initial signal investigations, at the end of 2004. Against this reference of emerging safety signals, the pattern is reversed and disproportionality analysis performs significantly better than raw numbers of reports. This lends empirical support to our previous cautionary note concerning performance evaluation of signal detection methods against established adverse drug reactions [15]: such evaluations should ideally be avoided or else interpreted with great caution. Furthermore, these results suggest that any comparison of analysis methods for individual case reports should include report counts as a comparator.

The sensitivity of spontaneous reporting rates to publication and selection biases is well-known, but the

[1] MedDRA®, the *Medical Dictionary for Regulatory Activities*, terminology developed under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). MedDRA® trademark is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA) on behalf of ICH.

[2] http://www.imi-protect.eu/documents/FinalRepository_DLP30Jun2012.xls

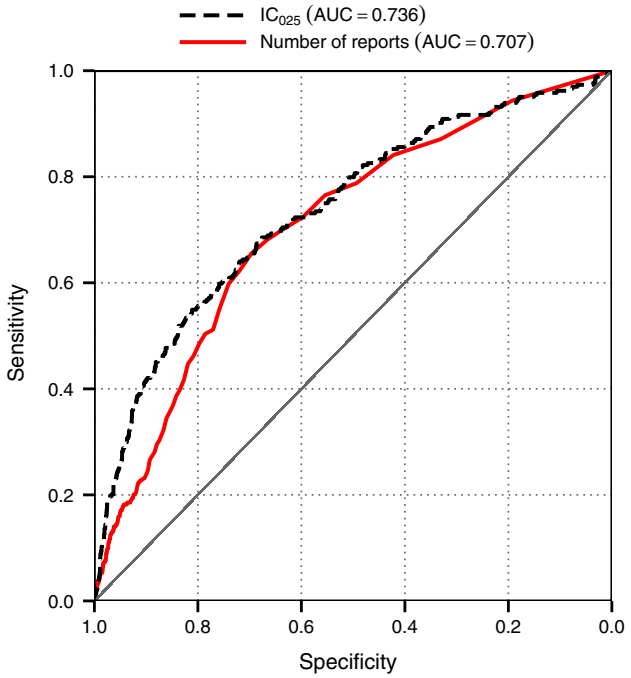

Fig. 2 Sensitivity and specificity for emerging adverse drug reactions of disproportionality analysis (IC_{025}) and raw report counts, respectively. The 264 positive controls of this reference set are pairs of drugs and MedDRA® PTs corresponding to historical safety signals derived from the study by Alvarez et al [14] backdated to around the time of the initial signal investigations (2004); the 5,280 negative controls are drugs paired with PTs for which no other PT in the same MedDRA® High-Level Term were listed on the drug's European SmPC in 2012. Data for both positive and negative controls were derived from a version of VigiBase® backdated to 2004. Thresholds for the disproportionality measure yield better or equal specificity than thresholds for the report count with the same sensitivity. The AUC values are 0.736 for IC_{025} and 0.707 for the raw report counts ($p < 0.05$ according to DeLong's test). *AUC* area under the receiver operating characteristic curve, IC_{025} lower limit of a 95 % credibility interval for the Information Component, *PT* Preferred Term, *SmPC* summary of product characteristics

distinction between established adverse drug reactions and emerging safety signals is also important for empirical evaluation of methods for screening longitudinal health records. Patient management will differ depending on whether an adverse event is believed to be causally associated with the treatment of interest, and this can have fundamental repercussions. As an example, a history of gastrointestinal bleeding can be expected to reduce the likelihood of future exposure to naproxen, as illustrated by the analysis of UK electronic patient records from The Health Improvement Network (THIN) shown in Fig. 3: upper gastrointestinal bleeding is overall less common in patients that receive naproxen, and particularly so in the months leading up to first naproxen prescriptions. Such explicit or implicit contraindications can make the risk more difficult to detect with cohort designs and will increase the apparent strength of association in self-controlled analyses due to the artificially low rate of the adverse event prior to first prescriptions. Taken together, these two effects will bias methodological comparisons for longitudinal health data in favour of self-controlled designs [16, 17]. Similarly, known risks of adverse drug reactions may stimulate closer monitoring for that adverse event under a particular treatment, as exemplified by the raised rate of acute liver injury (in this case, primarily abnormal liver function tests) on the day of first simvastatin prescriptions in Fig. 4. This is likely to reflect an increased rate of testing in these patients, since statins (HMG-CoA reductase inhibitors) are known to carry this risk. However, such patterns of intensified monitoring in direct conjunction with exposure are unlikely to occur for drugs not yet suspected to cause the adverse reaction, and so should not drive our choice of method for signal detection.

3 Related Work

While still in the minority, there are studies that have gone against the grain and evaluated methods against emerging safety signals. One interesting example was the evaluation undertaken by Bailey et al. [18], where prospective safety signals identified through regular pharmacovigilance activities during the course of the study were used as positive controls. This closely mimics the real pharmacovigilance setting and avoids the use of established adverse drug reactions for positive controls. However, a main limitation is the long time typically required to establish the true status of positive and negative controls, and their tentative nature along the way. Another significant challenge is that the signal detection activities to be evaluated affect the classification of positive and negative controls in non-trivial ways [19]. A more common approach has been to use historical safety signals as positive controls, backdating the data to before their initial identification, as in Fig. 2. An early example of this was the retrospective analysis of VigiBase® by Lindquist et al. [20], whereas more recent examples include the studies by Alvarez et al. [14] and Strandell et al. [21]. The reference set proposed by Alvarez et al. [14] is particularly interesting in that it provides dates not just for the regulatory action associated with each signal, but also the first dates that each signal was first discussed by the EMA's signal management team. Retrospective analyses are not suitable for evaluation of manual or semi-manual approaches since experienced safety scientists cannot be blinded to the true status of historical safety signals. However, beyond that, they are likely to be our best bet. A significant limitation of previous reference sets of emerging safety signals is their limited scope. An important step to improve the situation is the

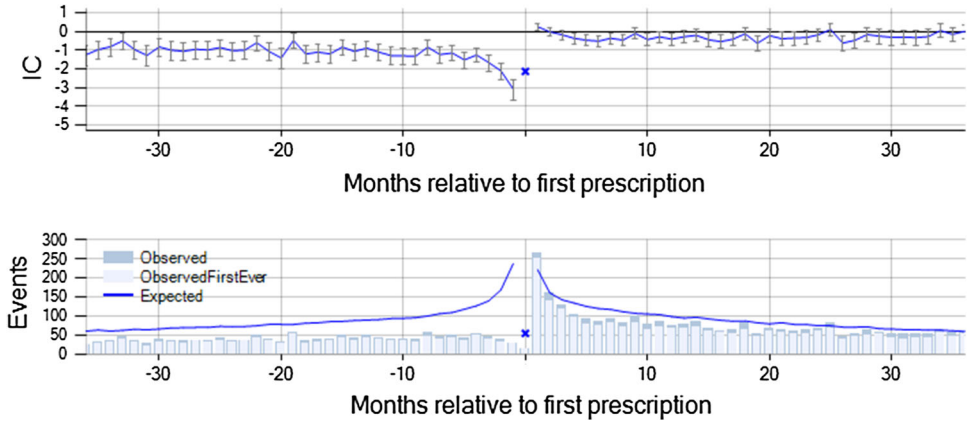

Fig. 3 Chronograph displaying the temporal pattern of upper GI bleeding events relative to first prescriptions of naproxen in THIN. There is an overall lower rate of upper GI bleeding events in patients prescribed naproxen, which is most pronounced in the months immediately prior to first naproxen prescription. The *x-axis* marks 30-day periods relative to first prescriptions of the drug (with the exception of time zero, which represents the day of prescription). The *bars* in the *bottom panel* represent the number of patients with a recorded upper GI bleeding event in each timeframe (with the number of patients who experienced their first such event ever in this time period marked in lighter shade), and the *line* indicates the corresponding expected values, which are based on the number of naproxen patients at risk and the rate of upper GI bleeding events at different times relative to other first prescriptions, in an external control group [2, 17]. The *upper panel* displays the base 2 logarithm of a shrinkage observed-to-expected ratio ('IC') with 95 % credibility intervals [2, 17]. THIN is a longitudinal observational health data from general practitioners in the UK. Upper GI bleeding events were ascertained based on 47 different READ codes, out of which J680.00 Haematemesis, J681.00 Melaena and J68z.11 GIB–Gastrointestinal bleeding were the most commonly used. *GI* gastrointestinal, *IC* Information Component, *THIN* The Health Improvement Network

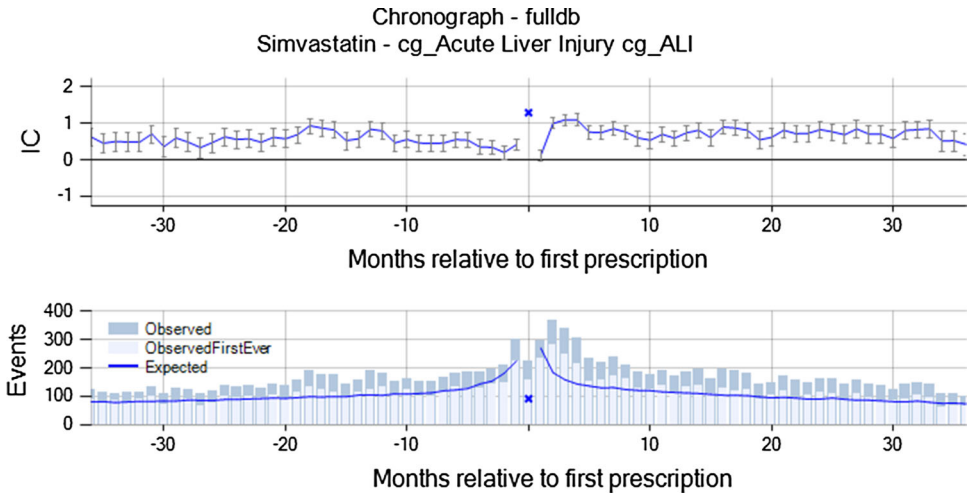

Fig. 4 Chronograph displaying the temporal pattern of acute liver injury events (in this instance, primarily reflecting abnormal liver function test values), relative to first prescriptions of naproxen in THIN. There is an increased rate of acute liver injury events on the day of first simvastatin prescriptions as well as after 2 months on simvastatin in the THIN database. The *x-axis* marks 30-day periods relative to first prescriptions of the drug (with the exception of time zero, which represents the day of prescription). The *bars* in the *bottom panel* represent the number of patients with a recorded acute liver injury event in each timeframe (with the number of patients who experienced their first such event ever in this time period marked in lighter shade), and the *line* indicates the corresponding expected values, which are based on the number of simvastatin patients at risk and the rate of acute liver injury events at different times relative to other first prescriptions, in an external control group [2, 17]. The *upper panel* displays the base 2 logarithm of a shrinkage observed-to-expected ratio ('IC') with 95 % credibility intervals [2, 17]. THIN is a longitudinal observational health data from general practitioners in the UK. Acute liver injury events were ascertained based on 58 different READ codes, out of which R148.11 LFT's Abnormal, 44D2.00 Liver function tests abnormal and R024.00 Jaundice (not of newborn) were the most commonly used. *IC* Information Component, *THIN* The Health Improvement Network

recent initiative to build an openly accessible knowledge base of all adverse drug reactions, which will include a time-stamp for every piece of evidence [22]; this will allow us to backdate our analyses to before adverse drug reactions were known, on a much grander scale than ever before.

4 Conclusions

The establishment of relevant reference sets of emerging safety signals must be made a top priority to achieve more effective pharmacovigilance methods development and evaluation. If done right, this might bring about just the type of savannah that we need: pharmacovigilance zebras dwelling in their natural habitat, challenging but not impossible to detect in the high grass. Such a training ground will help us discern which methods and information sources are most likely to bring value to prospective real-world surveillance for new adverse effects from drugs.

Acknowledgments None of the authors have received external funding to conduct this research. However, the example in Fig. 4 was identified during the course of a study within the PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium; http://www.imi-protect.eu) consortium, which has received support from the Innovative Medicine Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement n° 115004. G. Niklas Norén, Ola Caster, Kristina Juhlin and Marie Lindquist have no conflicts of interest that are directly relevant to the content of this study.

References

1. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18(6):427–36.
2. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov. 2010;20(3):361–87.
3. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc. 2013;20(3):404–8.
4. Freifeld CC, Brownstein JS, Menone CM, et al. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf. 2014;37(5):343–50.
5. Caster O, Norén GN, Madigan D, Bate A. Large-scale regression-based pattern discovery: the example of screening the WHO global drug safety database. Stat Anal Data Min. 2010;3(4):197–208.
6. 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.
7. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009;20(4):512–22 510.1097/EDE.1090b1013e3181a1663 cc.
8. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med. 2012;4(125):125–31.
9. Farrington CP, Nash J, Miller E. Case series analysis of adverse reactions to vaccines: a comparative evaluation. Am J Epidemiol. 1996;143(11):1165–73.
10. Ryan P, Schuemie M, Welebob E, Duke J, Valentine S, Hartzema A. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013;36(1):33–47.
11. Coloma P, 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.
12. Medical Products Agency (Läkemedelsverket). Läkemedelsverket utreder rapporter om narkolepsi efter vaccination med Pandemrix. 2010. http://www.lakemedelsverket.se/Alla-nyheter/NYHETER-2010/Lakemedelsverket-utreder-rapporter-om-narkolepsi-efter-vaccination-med-Pandemrix/. Accessed 11 Jun 2014.
13. Norén GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res. 2013;22(1):57–69.
14. Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in Eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. Drug Saf. 2010;33(6):475–87.
15. Caster O, Norén GN, Madigan D, Bate A. Logistic regression in signal detection: another piece added to the puzzle. Clin Pharmacol Ther. 2013;94(3):312.
16. Hallas J. Evidence of depression provoked by cardiovascular medication: a prescription sequence symmetry analysis. Epidemiology. 1996;7:478–84.
17. Norén GN, Bergvall T, Ryan PB, Juhlin K, Schuemie MJ, Madigan D. Empirical performance of the calibrated self-controlled cohort analysis within temporal pattern discovery: lessons for developing a risk identification and analysis system. Drug Saf. 2013;36(1):107–21.
18. Bailey S, Singh A, Azadian R, Huber P, Blum M. Prospective data mining of six products in the US FDA adverse event reporting system. Drug Saf. 2010;33(2):139–146.
19. Hauben M, Norén GN. A decade of data mining and still counting. Drug Saf. 2010;33(7):527–34.
20. Lindquist M, Stahl M, Bate A, Edwards IR, Meyboom RH. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf. 2000;23(6):533–42.
21. Strandell J, Caster O, Hopstadius J, Edwards IR, Norén GN. The development and evaluation of triage algorithms for early discovery of adverse drug interactions. Drug Saf. 2013;36(5):371–88.
22. Boyce RD, Ryan PB, Norén GN, et al. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Saf. doi:10.1007/s40264-014-0189-0.