Design criteria for reference sets in pharmacovigilance – The case of drug-drug interactions

Elpida Kontsioti, Simon Maskell, Munir Pirmohamed

Background

The challenge of building appropriate reference sets for performance evaluation of signal detection algorithms (SDAs) in pharmacovigilance has been widely acknowledged.[1–3] Previous studies have attempted to comparatively assess the performance of SDAs by generating custom-made reference sets, often limited in size, by applying specific inclusion and exclusion design criteria. Examples include those related to adverse event (AE) background prevalence, disease-related AEs, AE seriousness or evidence associated with positive controls.[4–7] This happens partially in an effort to address the limitations of disproportionality analysis, but also because observational data (e.g. postmarketing surveillance databases) suffer from confounding effects, in both directions, as a result of some variables remaining unobserved.[8] The presence of any synthetic associations (i.e. causative covariates that have been not taken into account and generates faulty associations between the drug and the AE), such as underlying conditions or concomitant medications, might be a source of selection bias and complicate the detection of safety signals.[9,10] Each SDA, depending on the applied modelling, might be impacted to a different extent by a confounder. Hence, the performance evaluation might be biased based on the selected benchmarks. We therefore need to consider the selection of appropriate controls to avoid misinterpretation of signals triggered by confounding factors rather than true associations as well as added biases to our evaluation by "favouring" some algorithms while penalising others.

Only limited efforts exist in the literature to generate reference sets related to two-way drug-drug interactions (DDIs).[11] A previous study has suggested that detection of DDI-related signals might suffer from multiple confounders.[12] Thus, we were interested in exploring the relative impact on the performance evaluation of three existing SDAs for DDI postmarketing surveillance when considering design criteria that could be applied to create reference sets in this setting.

Methods

Three SDAs previously described in the literature were considered: *Omega*, *delta_add* and *Interaction Signal Score* (*IntSS*).[13–15] A reference set of 4,455 positive and 4,544 negative controls for two-way DDIs was created by extracting and aggregating information from multiple clinical resources, namely the British National Formulary (BNF), the National Reference Guidance for healthcare professionals on DDIs (*Thesaurus des Interactions Médicamenteuses*) published by the French Medicines Agency (ANSM), and Micromedex. This reference set was the primary source of controls, covering 454 drugs and 179 adverse events mapped to OHDSI concepts (RxNorm and MedDRA vocabularies, respectively). A curated version of the FDA Adverse Event Reporting System (FAERS) database was used to generate evidence.[16]

We selected 13 design criteria that could be categorised as follows:

(1) Evidence level (only applied to positive controls)
    a. *BNF – Study* (for interactions where the information is based on formal study including those for other drugs with same mechanism, e.g. known inducers, inhibitors, or substrates of cytochrome P450 isoenzymes or P-glycoprotein)

b. *BNF – Theoretical* (interactions that are predicted based on sound theoretical considerations. The information may have been derived from *in vitro* studies or based on the way other members in the same class act)
c. *BNF – Anecdotal* (interactions based on either a single case report or a limited number of case reports)
d. *Micromedex – Established* (controlled studies have clearly established the existence of the interaction)
e. *Micromedex – Theoretical* (available documentation is poor, but pharmacologic considerations lead clinicians to suspect the interaction exists; or documentation is good for a pharmacologically similar drug)
f. *Micromedex – Probable* (documentation strongly suggests the interactions exists, but well-controlled studies are lacking)

(2) Event seriousness
   a. *EMA Important Medical Event (IME) Terms*
   b. *EMA Designated Medical Event (DME) Terms*

(3) Event frequency
   a. *Common AEs* (i.e. AE prevalence ≥ 90th percentile of prevalence of events reported in FAERS)
   b. *Rare AEs* (i.e. AE prevalence ≤ 10th percentile of prevalence of events reported in FAERS)

(4) *Potential confounding by indication* (i.e. the AE is also an indication for at least one of the two drugs from the drug-drug-event triplet under consideration)

(5) Potential confounding by concomitant medication
   a. *Shared indications – False* (i.e. drug pairs that share at least one indication are excluded)
   b. *Shared indications – True* (i.e. only drug pairs that share at least one indication are considered)

The positive and negative controls were stratified based on each of the above design criteria, forming suitable restricted subsets of different sizes in each case, depending on the criterion under consideration.

We simulated the generation of reference sets of multiple sizes ranging from 100 to N_{max} , where N_{max} was determined by the smaller of the two restricted sets (either the positive or the negative one) for each specific criterion. We randomly chose an equal number of positive and negative controls, either applying each of the design criteria mentioned above (restricted reference set, **RC1**) or not (unrestricted reference set, **RC2**) and calculated the Area Under the Curve (AUC) in both cases. The simulation was repeated 1000 times and AUC scores with 95% confidence intervals were calculated for both types. The difference of AUC scores (AUC_{diff}) was the target measure. The probability of AUC_{diff} being non-zero, $\mathrm{P}(|AUC_{diff}| > 0)$, was also estimated under the normality assumption. Figure 1 illustrates the framework for measuring differences in AUC scores for the different design criteria under consideration.

Results

Figure 2 shows the AUC_{diff} values for the different design criteria, SDAs, and reference set sizes. The size of the reference set did not have a considerable effect on AUC_{diff} of the different SDAs, which was the target metric. However, the associated probability of that metric being non-zero increased when considering larger sizes. By plotting AUC_{diff} for a fixed reference set size of 200 and ordering design

criteria by increasing range of AUC_{diff} values among the three SDAs (Figure 3), we can observe that some design criteria affected performance evaluation of all three SDAs in a similar way and level of magnitude

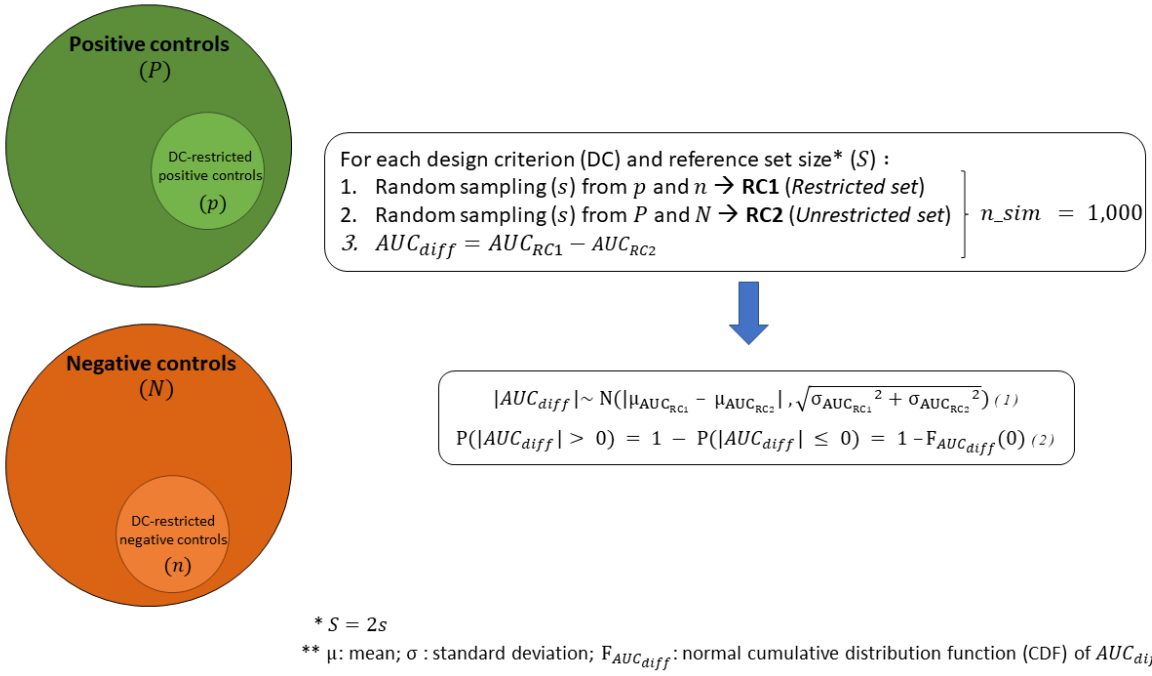

Figure 1. Testing framework for measuring differences in AUC scores for each design criterion.

Figure 2. AUC_{diff} values for the different design criteria, SDAs, and reference set sizes under consideration.

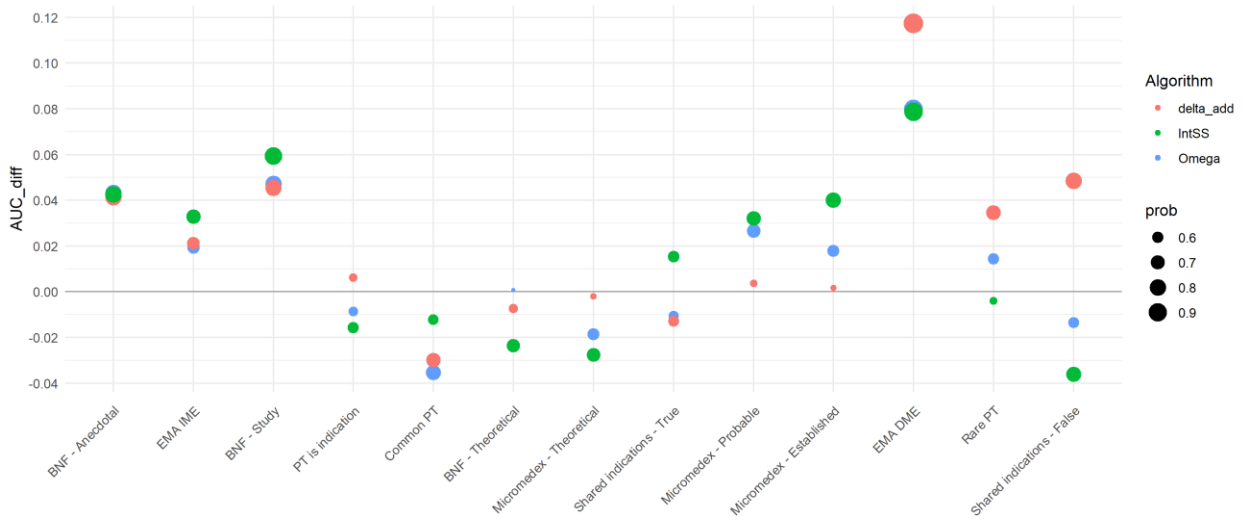

Figure 3. Ordered design criteria by increasing range of AUC_{diff} values among SDAs.

(e.g. BNF – Anecdotal), while others (e.g Shared indication – False) seemed to have opposing and different in size effects on AUC estimates. We also identified three main categories:

(i) Positive AUC_{diff} values
   a. BNF – Anecdotal
   b. EMA IME Terms
   c. BNF – Study
   d. Micromedex – Probable
   e. Micromedex – Established
   f. EMA DME Terms

(ii) Negative AUC_{diff} values
   a. Common AEs
   b. BNF – Theoretical
   c. Micromedex – Theoretical

(iii) Mixed effect on AUC_{diff} values
   a. AE is also an indication for at least one of the two drugs
   b. Only drug pairs that share at least one indication are included
   c. Rare AEs
   d. Drug pairs that share at least one indication are excluded

Conclusion

This study revealed a varying impact of design criteria for reference sets on the AUC scores of three SDAs that are used for DDI postmarketing surveillance. This analysis showcases that the design of reference sets should be performed carefully, as the comparison of SDA performance might be affected by the choices made when building a reference set and the decision to restrict the evaluation to specific controls.

Also, it highlights the need for establishment of open and sizable benchmarks that include a diverse set of controls to ensure transparency and enable a fair evaluation of SDA performance.

References

1. Boyce RD, Ryan PB, Norén GN, Schuemie MJ, Reich C, Duke J, 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.
2. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance. Vol. 37, Drug Safety. Springer International Publishing; 2014. p. 655–9.
3. Harpaz R, DuMouchel W, Shah NH. Comment on: "Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance." Vol. 38, Drug Safety. 2015. p. 113–4.
4. 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).
5. Hoffman KB, Dimbil M, Tatonetti NP, Kyle RF. A Pharmacovigilance Signaling System Based on FDA Regulatory Action and Post-Marketing Adverse Event Reports. Drug Saf. 2016;39(6):561–75.
6. Seabroke S, Candore G, Juhlin K, Quarcoo N, Wisniewski A, Arani R, et al. Performance of Stratified and Subgrouped Disproportionality Analyses in Spontaneous Databases. Drug Saf. 2016;39(4):355–64.
7. Arnaud M, Bégaud B, Thiessard F, Jarrion Q, Bezin J, Pariente A, et al. An Automated System Combining Safety Signal Detection and Prioritization from Healthcare Databases: A Pilot Study. Drug Saf. 2018;41(4):377–87.
8. Hopstadius J, Norén GN, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. Drug Saf. 2008;31(11):1035–48.
9. Tatonetti N, Ye P, Daneshjou R, Altman R. Data-Driven Prediction of Drug Effects and Interactions. Sci Transl Med. 2012;4(125):125ra31-125ra31.
10. Dijkstra L, Garling M, Foraita R, Pigeot I. Adverse drug reaction or innocent bystander? A systematic comparison of statistical discovery methods for spontaneous reporting systems. Pharmacoepidemiol Drug Saf. 2020;29(4):396–403.
11. Juhlin K, Soeria-Atmadja D, Thakrar B, Norén GN. Evaluation of statistical measures for adverse drug interaction surveillance. Pharmacoepidemiol Drug Saf Drug Saf. 2014;23(S1):294–5.
12. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics. 2010 Oct 28;11(S9):S7.
13. Norén GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug–drug interaction surveillance. Stat Med. 2008;27:3057–70.
14. Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug-drug interactions in a spontaneous reports database. Br J Clin Pharmacol. 2007 Oct;64(4):489–95.
15. Almenoff JS, DuMouchel W, Kindman LA, Yang X, Fram D. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidemiol Drug Saf. 2003 Sep 1;12(6):517–21.
16. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. Data Descriptor: A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016;3.