Kontsioti Elpida (Orcid ID: 0000-0002-4053-0220)
Pirmohamed Munir (Orcid ID: 0000-0002-7534-7266)

Exploring the impact of design criteria for reference sets on performance evaluation of signal detection algorithms: the case of drug-drug interactions

Running heading

Design criteria for reference sets in signal detection of drug-drug interactions

Author information

Elpida Kontsioti [a,*], Simon Maskell [a], Munir Pirmohamed [b]

a Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom

b The Wolfson Centre for Personalized Medicine, Centre for Drug Safety Science, Department of Pharmacology and Therapeutics, Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom

*Corresponding author (E.Kontsioti@liverpool.ac.uk)

ORCID IDs

Elpida Kontsioti (0000-0002-4053-0220)

Simon Maskell (0000-0003-1917-2913)

Munir Pirmohamed (0000-0002-7534-7266)

Keywords

spontaneous reports data; signal detection; performance metrics; pharmacovigilance; postmarketing surveillance; adverse events; drug-drug interactions

Abstract

Purpose

To evaluate the impact of multiple design criteria for reference sets that are used to quantitatively assess the performance of pharmacovigilance signal detection algorithms (SDAs) for drug-drug interactions (DDIs).

Methods

Starting from a large and diversified reference set for two-way DDIs, we generated custom-made reference sets of various sizes considering multiple design criteria (e.g., adverse event background prevalence). We assessed differences observed in the performance metrics of three different SDAs when applied to FDA Adverse Event Reporting System (FAERS) data.

Results

For some design criteria, the impact on the performance metrics was neglectable for the different SDAs (e.g., theoretical evidence associated with positive controls), while others (e.g., restriction to designated medical events, event background prevalence) seemed to have opposing and effects of different sizes on AUC and PPV estimates.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pds.5609

Conclusions
The relative composition of reference sets can significantly impact the evaluation metrics, potentially altering the conclusions regarding which methodologies are perceived to perform best. We therefore need to carefully consider the selection of controls to avoid misinterpretation of signals triggered by confounding factors rather than true associations as well as adding biases to our evaluation by "favouring" some algorithms while penalising others.

Key points

- Performance assessment of SDAs in pharmacovigilance has often relied on the generation of custom-made reference sets of limited size that consider ad-hoc exclusion or inclusion criteria to define eligible controls.
- SDA performance assessment might be biased based on the selected benchmarks, as each methodology can be impacted to a different extent by different confounders.
- We tested 14 design criteria for reference sets in the case of DDIs, showing that some of them considerably affected the performance and comparative evaluation of different SDAs for DDI surveillance while others did not have a significant effect.
- Overall, this analysis advocates the utilisation of large, to the extent possible, reference sets that are less likely to suffer from overrepresentation of controls that make different SDAs behave in different ways due to confounding. Any decision to restrict the evaluation set using specific design criteria should be carefully justified.

Plain Language Summary

Reporting of suspected side effects experienced by patients following drug approval is a key component to identify novel drug safety issues. Statistical methods are then used to analyse reports and reveal signals of novel associations between drugs and side effects. Performance evaluation of those methods traditionally relies on custom-made reference sets of limited size that consider ad-hoc exclusion or inclusion criteria to define eligible controls. However, each method can be impacted to a different extent by those criteria, as they can act as potential confounders. This study investigated the impact of 14 criteria on three methods that have been developed to detect signals of potential adverse drug-drug interactions, showing that some of them had opposing effects or effects of different levels of magnitude on the performance of the different methods. The relative composition of reference sets can therefore significantly affect the evaluation metrics, potentially altering the conclusions regarding which methodologies are perceived to perform best. The selection of controls should be carefully performed to avoid misinterpretation of signals triggered by confounding factors rather than true associations as well as adding biases to our evaluation by "favouring" some algorithms while penalising others.

Acknowledgements

The authors would like to thank Dr Bhaskar Dutta, Dr Isobel Anderson, and Mr Antoni Wisniewski for their input and fruitful discussions.

Declarations

Funding

This study was jointly funded by EPSRC (grant number EP/R51231X/1) and AstraZeneca.

Conflicts of interest

Elpida Kontsioti received PhD studentship that was jointly funded by AstraZeneca and the EPSRC. She is currently an employee of The Hyve BV. Munir Pirmohamed receives research funding from various organizations including the MRC and NIHR. He has also received partnership funding for the MRC Clinical Pharmacology Training Scheme (co-funded by MRC and Roche, UCB, Eli Lilly and Novartis) and grant funding from Vistagen Therapeutics. He has also unrestricted educational grant support for the UK Pharmacogenetics and Stratified Medicine Network from Bristol-Myers Squibb and UCB. He has developed an HLA genotyping panel with MC Diagnostics, but does not benefit financially from this. He is part of the IMI Consortium ARDAT (www.ardat.org). These funding sources were not utilized for this work. Simon Maskell declares that he has no conflict of interest.

Data availability statement

The CRESCENDDI data set that supports the findings of this study is openly available in Figshare at https://doi.org/10.6084/m9.figshare.c.5481408.v1.

Authors' contributions

EK, SM, and MP contributed to the study conception and design. Material preparation, data collection and analysis were performed by EK. The first draft of the manuscript was written by EK and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

1 Introduction

Monitoring drug safety issues during the post-approval phase requires reporting of suspected drug-related adverse reactions by healthcare professionals, patients, and pharmaceutical companies. The reports are collected in spontaneous reporting system (SRS) databases, such as the FDA Adverse Event Reporting System (FAERS) database in the US, the Eudravigilance database in the EU, and Yellow card database in the UK. These databases form an important part of the pharmacovigilance strategy since they do not only contain information on adverse events (AEs) and suspected drugs, but also details regarding concomitant medications, indications, and patient demographics.

By applying statistical methods known as signal detection algorithms (SDAs), novel associations between drugs and AEs (i.e., signals) that have not been identified in clinical trials can be identified in the SRS data. Given the absence of a control group, SDAs predominantly rely on disproportionality analysis, which calculates the degree of disproportional reporting of drug-AE combinations compared to what would be expected if there was no association between them.[1] However, the presence of synthetic associations (i.e., causative covariates that have not been taken into account or remain unobserved) can lead to confounding, either upward or downward, thus generating faulty associations between the drug and the AE and complicating the detection of safety signals.[2–4] For example, reporting quality issues arising from a poor distinction between symptoms of disease-related AEs and treatment effects of drugs (or drug combinations) is a result of a synthetic association called *confounding by indication*.[5,6]

The practice of using larger clusters of medical terms to perform quantitative signal detection in pharmacovigilance has been widely discussed in the literature.[1,7] Many previous efforts investigated the impact of MedDRA granularity on signal detection tasks.[8,9] Also, many studies have considered the use of term grouping to identify relevant reports.[10,11] However, recommendations from the IMI-PROTECT project suggest that signal detection at the PT level should be considered the standard approach in real-life pharmacovigilance.[9,12]

The development of novel SDAs in pharmacovigilance requires the existence of appropriate reference sets that can be utilized both for absolute performance evaluation as well as for comparison with existing methodologies. Given that each SDA, depending on the applied modelling, might be impacted to a different extent by a confounder, the performance evaluation might be biased based on the selected benchmarks. The challenge of building appropriate reference sets in pharmacovigilance has been previously acknowledged in the literature.[13–16] Most studies have attempted to comparatively evaluate SDAs by testing their performance against custom-made reference sets, often limited in size[17–19] or not publicly available[20,21] , which commonly consider ad-hoc inclusion or exclusion criteria to generate positive and negative controls. Examples of such criteria include those related to AE background prevalence (given that, in disproportionality analysis, the denominator signifies the expected rate of occurrence)[22], disease-related AEs[23], AE seriousness[23,24] or evidence associated with positive controls[22–26]. The criteria are typically used to attempt to address the limitations of disproportionality analysis and to tackle issues with potential confounders.

In the case of adverse drug-drug interactions (DDIs), signal detection is considered more complicated, with the existing methodology being less mature compared to the one in the case of signals for single drugs. A previous study has suggested that detection of DDI-related

signals might suffer from multiple confounders.[27] For example, concomitant medications appear to be a significant source of confounding (i.e., the signal associated with a drug combination was triggered by drugs are usually given concomitantly but not signify true adverse drug-drug-event associations). In addition, only limited efforts exist in the literature to generate reference sets related to two-way DDIs.[17,19,27,28]

In this study, we aim to explore the relative impact of different factors that could be potential sources of confounding on the performance evaluation of existing methods for signal detection of DDIs. By utilising a large and diversified reference set, we were able to create custom-made reference sets considering multiple design criteria to assess any differences observed in the quantitative evaluation of SDAs tailored for two-way DDIs.

2 Methods

2.1 Data Sources

2.1.1 FAERS data – Spontaneous reports

We used a curated and standardized version of the publicly available FAERS database. The data pre-processing pipeline was based on the Adverse Event Open Learning through Universal Standardization (AEOLUS) process and included removal of duplicate reports, drug name normalization at the RxNorm ingredient level, and AE mapping to MedDRA Preferred Terms (PTs).[29] The curated data set included 9,203,239 reports containing at least one drug and one AE between 2004 (Q1) and 2018 (Q4), with 3,973,749 (43.18%) reports mentioning more than one drug. Each drug was considered equivalent in the analysis irrespective of its reported role (i.e., *primary suspect*; *secondary suspect*; *concomitant*; and *interacting*).

2.1.2 Reference sets for DDIs

CRESCENDDI, a reference set for two-way DDIs, was the primary source of controls.[30] This reference sets covers 454 drugs and 179 adverse events mapped to RxNorm Ingredient and MedDRA PT concepts, respectively, from the Observational Medical Outcomes Partnership (OMOP) Common Data Model (version 5). We used 4,455 positive and 4,544 negative controls from CRESCENDDI that were also present in the curated FAERS dataset (hereafter called ***PT Reference Set***).

To accommodate and test the impact of MedDRA granularity to detect signals at the medical concept (MC) level, we extended CRESCENDDI by building PT groups (*event groups*), where possible, that are relevant to the adverse events described in the original reference set. These groups were formed by examining Standardised MedDRA Queries (SMQs) and event definitions from a time-indexed reference standard by Harpaz et al.[31] and were manually reviewed for clinical relevance. In total, 20 adverse events from CRESCENDDI deemed suitable for extension to the MC level (**Table 1**). A full list of the event groups is available in **Appendix S1**. The new reference set (hereafter called ***MC Reference Set***) contained 1,097 positive and 614 negative controls (**Appendix S2**).

2.2 Data mining

We performed the case/non-case analysis at two different levels, based on the reference sets that we utilised. The first one was restricted to the reports that included the PT that was

related to each control from the ***PT Reference Set***. The second one considered as cases all the reports that contained any of the PTs that were part of the MC linked to the control in the ***MC Reference Set***.

For example, the case/non-case analysis for a control related to torsade de pointes resulted in two contingency tables: the first one only considered the PT '*Torsade de pointes*' to retrieve case reports, while the second one included the following terms (as PTs): '*Electrocardiogram QT interval abnormal*', '*Electrocardiogram QT prolonged*', '*Long QT syndrome*', '*Torsade de pointes*', '*Ventricular tachycardia'*. Non-cases included the reports without the aforementioned PTs, while reports containing more than one of the relevant PTs linked to the MC were not double counted.

2.3 Design Criteria

Table 2 shows the design criteria that were considered as potential confounding factors, which fall into the following categories: (i) evidence level; (ii) event seriousness; (iii) event frequency; (iv) potential confounding by indication; and (v) potential confounding by concomitant medication. ***PT Reference Set*** controls were stratified based on each of the design criteria, forming suitable restricted subsets of different sizes in each case, depending on the criterion under consideration. ***MC Reference Set*** could not be stratified using categories (ii) and (iii).

2.4 PT prevalence

The impact of reference set restriction by PT prevalence on AUC estimates was also examined. The PT prevalence was calculated in the curated FAERS data set as the frequency of PTs from reports containing at least one drug. We grouped the 179 PTs from the ***PT Reference Set*** using quartile binning of their prevalence. The controls were then stratified in 4 groups (Groups Q1-Q4) based on their PTs by considering the respective PT prevalence quartile.

2.5 SDAs

Three SDAs that have been previously described in the literature were considered:

(i) An observed-to-expected shrunk interaction measure (*Omega*)[32];
(ii) The 'interaction coefficient' in a linear regression model with additive baseline (*delta_add*)[33];
(iii) A measure based on an adapted version of Multi-Gamma Poisson Shrinker (MGPS) model, called *Interaction Signal Score* (*IntSS*)[17].

2.6 Impact of MedDRA Granularity on SDA Performance Evaluation

To assess the impact of MedDRA granularity on the SDAs that were considered in this study, we performed a Receiver Operating Characteristic (ROC) analysis to examine the difference in the Area Under the Curve (AUC) when considering matched controls from the two reference sets.

2.7 Estimation of Design Criteria Impact on SDA Performance Evaluation

For each reference set and design criterion, we simulated the generation of a *constrained reference set* by randomly drawing an equal number (1:1) of positive and negative controls from the restricted control subset that used the specified design criterion for control stratification. An *unconstrained reference set* of equal size was generated in each case by following a similar process but using the original reference set.. This sampling generation process took into account the correlation between the two sets, as the probability of drawing one control for the *constrained reference set* did not affect the probability of drawing any control for the *unconstrained reference set*. The size of the simulated reference sets varied from 100 to $2 \times N_{max}$, where N_{max} was determined by either the number of positive or negative controls (depending on which one was smaller) in each of the restricted subsets. For each SDA, we calculated: (i) AUC scores; and (ii) PPV for fixed sensitivity values (i.e., 0.60, 0.75, and 0.90)for both reference set types (i.e., *constrained* and *unconstrained*) by performing 1,000 simulations. The statistics of the samples were summarised by fitting a Normal distribution, for which we report the mean and variance. The difference of the means of AUC (AUC_{diff}), and PPV (PPV_{diff}) (with 95% confidence intervals) were the target measures. The probability of AUC_{diff} being non-zero, $\mathrm{P}(|AUC_{diff}| > 0)$, was also estimated under the normality assumption:

$$|AUC_{diff}| \sim \mathrm{N}\left(|\mu_{\mathrm{AUC}_{Restricted\_ROC}} - \mu_{\mathrm{AUC}_{\mathrm{Unrestricted}\_ROC}}|, \sqrt{{\sigma_{\mathrm{AUC}_{Restricted\_ROC}}}^2 + {\sigma_{\mathrm{AUC}_{\mathrm{Unrestricted}\_ROC}}}^2}\right) \quad (1)$$

$$\mathrm{P}(|AUC_{diff}| > 0) = 1 - \mathrm{P}(|AUC_{diff}| = 0) = 1 - \mathrm{F}_{AUC_{diff}}(0) \quad (2)$$

where μ is the mean, σ is the standard deviation, and $\mathrm{F}_{AUC_{diff}}$ is the normal cumulative distribution function (CDF) of AUC_{diff}.

Figure 1 illustrates the simulation workflow for the calculation of differences in AUC scores and PPV when considering the various design criteria.

3 Results

The total number of positive and negative controls when applying each of the design criteria to the ***PT Reference Set*** is presented in **Figure 2**. In cases where restricted subsets contained both positive and negative controls (**Figure 2a**), the maximum number of controls considered from each type (i.e., positive or negative) to form simulated reference sets ($\boldsymbol{N_{max}}$) is denoted with white color in the respective bar. For the design criteria under the *Evidence level* category, where restriction was only applied to positive controls (**Figure 2b**), $\boldsymbol{N_{max}}$ was defined as the total number of positive controls in the respective restricted subsets. Apart from two cases (i.e., *Shared indications - False* and *AE is an indication – False*), positive controls outnumbered negative controls in the restricted subsets. The simulated reference sets varied in size, with N_{max} ranging from 131 to 3,568. Hence, more than 250 positive and negative controls were considered for every design criterion. For the ***MC Reference Set***, the restricted subsets were smaller in size (**Supplementary Table S1**). Three design criteria (*BNF – Anecdotal*, *BNF – Theoretical*, and *AE is an indication – True*) were not tested with this reference set, as their $\boldsymbol{N_{max}}$ was less than or equal to 100. **Figure 3** provides the frequency distribution of PT prevalence in: (a) the set of unique PTs in the ***PT Reference Set***; (b) ***PT***

Reference Set positive controls; and (c) ***PT Reference Set*** negative controls. The right-tailed distribution of unique PTs in CRESCENDDI shows that the data set was populated with less common PTs, with only small number of them having a prevalence over 0.01 in FAERS. Similar trends were present in the curves of the positive and negative controls, with the latter consisting of more cases with a higher PT prevalence in FAERS. The 1st, 2nd and 3rd quartiles for the PT prevalence were 0.000343, 0.00135, and 0.00410, respectively. The total number of positive and negative controls for each group formed using PT prevalence quartile binning is shown in **Figure 4**. *Group Q3* contained the largest volume in the case of positive controls, with *Group Q1* and *Group Q2* being considerably smaller, while negative controls showed an increasing trend while moving to groups of higher PT prevalence.

The MedDRA granularity affected the SDA performance metrics in different ways (**Table 3**). *Omega* and *IntSS* performed worse at the MC level as opposed to the PT level, with their mean AUC score dropping by 0.0605 and 0.0489, respectively. For *Omega*, here was a statistically significant decrease in the AUC between the PT and MC level evaluations. In the case of *delta_add*, the mean AUC slightly increased (0.0311) when considering the MC level, however without outperforming *Omega*.

By plotting AUC_{diff} for a fixed constrained reference set size of 100 and ordering design criteria by increasing range of AUC_{diff} values among the three SDAs **(Figures 5, S1**), points that lie above the x-axis signify positive estimates for AUC_{diff}, meaning that the design criterion had a positive effect on the calculated AUC. Conversely, points below the x-axis were associated with negative effect on the AUC when the specific design criterion was applied to constrain the reference set. Also, for the different sizes of restricted reference sets using the ***PT Reference Set*** and the ***MC Reference Set***, AUC_{diff} value estimates and associated probabilities of a non-zero AUC_{diff} estimate were plotted (**Figures S2, S3**). With the ***PT Reference Set***, the largest AUC_{diff} values were associated with the *EMA Designated Medical Event Terms* criterion (between 0.071 and 0.095), while *Common PTs* resulted in negative values in the range of -0.041 to -0.021 for the AUC_{diff} measure for all SDAs. In the case of the ***MC Reference Set***, *BNF – Study* had the largest positive impact on all AUC_{diff}values (between 0.098 and 0.051), while negative AUC_{diff}values derived from *Shared indications – True* and *AE is an indication – False* (up to -0.043). Some design criteria affected performance evaluation of all three SDAs in a similar way and level of magnitude (e.g., *BNF – Anecdotal*, *BNF – Study*), while others (e.g., *Shared indication – False*) seemed to have opposing and different in size effects on AUC estimates.

Supplementary Tables S2-S3 report the PPV_{diff} estimates (with 95% CIs) for the different design criteria, and a fixed reference set size of 100, for the ***PT Reference Set*** and ***MC Reference Set***, respectively. For both reference sets and a sensitivity equal to 0.60, some design criteria affected PPV in opposing ways among the different SDAs. For example, *Shared indications – False* resulted in negative PPV_{diff} estimates for *Omega* and *IntSS* (in the range between -0.029 and -0.021) as opposed to positive ones for *delta_add* (around 0.051). For other design criteria (i.e., *BNF – Study* and *EMA – Designated Medical Events*), PPV_{diff} estimates were positive across the different sensitivity values for all three SDAs. For a sensitivity value of 0.90, PPV_{diff} for the different design criteria were close to zero in all cases (values between 0.029 and -0.009).

With the ***PT Reference Set***, we identified three main categories:

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

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

(iii) Mixed effect on AUC_{diff} values
   a. AE is an indication - False
   b. AE is an indication - True
   c. Micromedex – Established
   d. BNF – Theoretical
   e. Only drug pairs that share at least one indication are included
   f. Drug pairs that share at least one indication are excluded

With the ***MC Reference Set*** study, *Omega* and *IntSS* were affected in a similar way by the different design criteria. *BNF – Study* and *Micromedex - Established* had a positive impact on the target measure for all SDAs, while excluding AEs related to drugs' indications (*AE is an indication – False*) or only considering drug pairs with shared indications as controls (*Shared indications – True*) negatively affected the SDA performance in all cases.

In terms of PT prevalence (**Figure 6**), there was a similar trend for *Groups Q1* to *Q3*, with AUC_{diff} metric increasing for all algorithms as we moved to more common PTs. However, this relationship appears to be reversed in *Group Q4*, which contains the most frequent PTs in FAERS from the original data set, for *Omega* and *delta_add*, showing a negative impact on their AUC.

4 Discussion

This study provides a systematic evaluation of the impact of multiple design criteria for reference sets on the comparative assessment of signal detection methodologies of adverse DDIs in SRS data. Performance assessment of SDAs in pharmacovigilance has often relied on the generation of custom-made reference sets that consider exclusion or inclusion criteria to define eligible controls. Thus, the motivation behind this research was to examine how different criteria could affect the evaluation, potentially altering the conclusions regarding which algorithms perform best.

Our study highlighted that the relative composition of reference sets might significantly impact the evaluation metrics. Some criteria affect the comparison of different methodologies, such as the restriction of controls to only include PTs from the EMA's designated medical event list. Other criteria that were thought to have a potential effect on the evaluation process (e.g., anecdotal evidence supporting a positive control) were not found to significantly change the observed difference in metrics amongst the methodologies, as all of them were influenced in a similar way (**Figure 5**). Moreover, we found that the size of the reference set did not have

a considerable effect on the AUC_{diff}, although the associated probability of that metric being non-zero increased when considering larger sizes (**Figures S1-S2**). Apart from AUC, commonly applied sensitivity values were considered to identify the impact of design criteria on PPV. For most of the design criteria (e.g., *EMA Designated Medical Events*, Micromedex evidence categories), PPV_{diff} values were affected consistently with the AUC_{diff} estimates across the three different SDAs. For the highest sensitivity that was considered (0.90), the difference in PPV was in most cases neglectable.

Given the inability of SDAs to account for all potential confounding factors that are present in SRS data, each methodology might be impacted to a different extent by a confounder. At the same time, there might be cases where signals are triggered by those confounding factors. As an illustrative example, the majority of DDI signals identified using *IntSS* in the original research paper[27] were composed of drug pairs that are usually given concomitantly (e.g., antibiotics).[27] 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 adding biases to our evaluation by "favouring" some algorithms while penalising others. On the other hand, by attempting to completely remove all potential sources of confounding in our evaluation sets, we are more likely to fail to demonstrate their utility in real-life application, which should be determined by its ability to perform at a commensurate level when it is applied prospectively to identify novel signals in SRS databases.[14,15] Overall, this analysis advocates the utilisation of large, to the extent possible, reference sets when it comes to comparative performance assessment, that are less likely to suffer from overrepresentation of controls that make different SDAs behave in different ways due to confounding. Also, regarding novel reference sets, the decision to restrict the evaluation set using specific design criteria should be adequately supported.

A major concern about reference sets used for prospective signal detection in pharmacovigilance revolves around the validity of established (i.e., well-known) positive controls to test the performance of algorithms. This aspect has been widely discussed in the literature.[14,15,34] It has been acknowledged that the combination of established and emerging positive controls might be a better choice when we try to evaluate the prospective performance and compare different methodologies, because merely emerging positive controls (i.e., recently detected ADRs) cannot establish a reliable reference standard.[18] Especially for DDIs, the establishment of reference sets by only using emerging positive controls turns out to be particularly challenging, as we would end up having a very limited number of controls to be able to quantitatively assess differences in the performance of the SDAs under comparison. A solution to this issue would be to perform a backdated analysis to detect the time point of that a signal of a true positive association ('positive control') was first highlighted, as proposed in previous studies.[35] However, this backdated analysis was not possible in this study due to the lack of a time-indexed reference set for DDIs. A previous study compared the performance of SDA algorithms for DDI surveillance between established and emerging positive controls, with *Omega* and *delta_add* showing increased specificity but diminished sensitivity in the latter case.[19] In our analysis, the results related to evidence level are consistent with what we would expect to see. In terms of *theoretical DDIs*, it is common for drug interaction compendia to extend the included DDIs to the drug class level, therefore covering drugs under the same drug class that sometimes, but not necessarily, have a similar interaction profile. Our results showed declining AUC when considering theoretical DDIs

(i.e., *Micromedex – Theoretical*) as opposed to improvements with established ones (i.e., *BNF – Study* and *Micromedex – Established*). On the other hand, all three examined methodologies demonstrated enhanced performance against *anecdotal DDIs* from BNF and *probable DDIs* from Micromedex. However, the former category represented only a small fraction of the overall positive cases contained in the ***PT Reference Set*** (2.94%).

In terms of event background prevalence, the simulation results suggest that, if we restricted the evaluation set to specific ranges of PT prevalence, the conclusions would change, i.e., the sole choice of common PTs would have an inverse impact on the comparative evaluation as to rare AEs. We know that SRS data are predominantly used in the post-marketing setting to spot rare adverse reactions that have not been revealed during clinical trials. However, the use of SRS data for the detection of DDIs can be considered a different scenario, given that clinical trial data are not sufficient to detect adverse reactions of drug combinations due to inherent limitations (e.g., patient recruitment processes that excludes people taking multiple medications). Hence, the detection of novel DDI-related adverse reactions, even with a common background rate, in SRS data should be of special interest.

Disease-related AEs are a challenging issue in the effort to generate signals using SRS data, as confounding by indication can occur. A previous study reported that around 5% of the total reports for any drug in FAERS mention a drug's indication as an adverse event.[36] This might be related to poor reporting quality or intended to report a disease's exacerbations due to a drug. Our results support that the choice of excluding disease-related AEs (i.e., *AE is an indication – False*) did not have a significant effect on the AUC across the SDAs with the ***PT Reference Set***, while it decreased the performance of all SDAs with the ***MC Reference Set***. On the other hand, *Omega* demonstrated deteriorated performance in the scenario of detecting controls with AEs that were drugs' indications at the same time (i.e., *AE is an indication – True*), while the other two SDAs did not seem to be substantially affected by this design criterion.

Event seriousness has been used to build reference sets and assess SDA performance, as it could be utilised to filter signals in real-life pharmacovigilance settings.[23,24] Our study suggests that, by only considering 'significant' events, bias is introduced to evaluating SDAs that could be potentially used in routine pharmacovigilance to detect a broader set of events. Also, given that DMEs are rare events (i.e., have low prevalence) with a high drug-attributable risk, it is important to note that this category might have been confounded to an extent by other design criteria categories that were considered in our study, such as the event frequency.

Quantitative signal detection is only one aspect of the more complex framework before a safety signal is validated. In the case of adverse DDI surveillance, previous studies have considered triage filters alongside disproportionality analysis to direct preliminary signal assessment.[37,38] These filters might be less suitable depending on the type of DDI. For example, there are more filters relevant to pharmacokinetic DDIs (e.g., cytochrome P450 activity) as opposed to pharmacodynamic interactions. Although the clinical significance of the differences between SDAs that are reported in this study might be questioned, it is important to note that quantitative methods for adverse DDI surveillance remain way less mature compared to those for single-drug safety surveillance, also considering the additional complexity that is inherent to DDIs. In this way, the potential impact on real-world

pharmacovigilance could not be refuted, as even small changes in the performance of an SDA might have a considerable impact on the number of generated signals that are captured for further evaluation, leading to either missed signals or large amounts of potential signals that need to be evaluated, thus increasing the manual effort needed. It is also important to note that the three SDAs that were included in our study are not implemented to the same extent in the real-world. *Omega* and *IntSS* are two of the major methods that we understand to be used for routine pharmacovigilance screening for DDIs. *delta_add* is a less mature method that is described in the literature, for which, as far as we are aware, is not as widely used in practice.

Although this study provides a novel framework for studying how SDA performance may change by considering different criteria for eligibility of controls, there are some limitations worth mentioning. First, only a single test data set (i.e., FAERS) was utilised for the purposes of this study. Also, CRESCENDDI was the only reference set utilised to generate estimates of the impact on AUC, in the absence of another comprehensive data set that could be used as a comparative source. We acknowledge that, by modifying the CRESCENDDI data set to consider adverse events at the MC level, we ended up with a smaller reference set that only included controls that could be represented by event groups (e.g., angioedema). This might have an impact on the extrapolation of the results and conclusions drawn from our analysis when considering single PTs as opposed to event groups. Additionally, for the determination of hit versus miss, it is important to consider how the results calculated at the PT level can depict the signal generation at the MC level. For example, if one SDA signals *polymorphous ventricular tachycardia* and another one signals *torsade des points* at the PT level, they have both made the same classification in real-world pharmacovigilance, as both would have triggered the same case review by a diligent pharmacovigilance organization. The performance of SDAs was only assessed using the default values provided in the original research papers describing those methods (e.g., tuning parameter for shrinkage, *a*, equal to 0.5 in the case of *Omega*). Finally, the aspect of unbalanced reference sets was not explored in this study (i.e., positive to negative control ratio different than 1:1), since previous studies in pharmacovigilance have evaluated SDAs using asymmetrical reference sets.[18,24,31]

5 Conclusions

This study revealed a varying impact of design criteria for reference sets on the performance metrics of three SDAs that are used for DDI post-marketing 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 to establish frameworks that can make use of large and disparate data sources to support the generation of open-source, flexible benchmarks in pharmacovigilance. These benchmarks can not only ensure transparency and enable a fair evaluation of SDA performance, but also provide a strong foundation that promotes productive research in pharmacovigilance signal detection methodologies.

References

1. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf* 2009; **18**: 427–436. doi:10.1002/pds.1742.

2. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects

and interactions. *Sci Transl Med* 2012; **4**. doi:10.1126/scitranslmed.3003377.

3. 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**: 396–403. doi:10.1002/pds.4970.

4. Hopstadius J, Norén GN, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. *Drug Saf* 2008; **31**: 1035–1048. doi:10.2165/00002018-200831110-00008.

5. Catalogue of bias collaboration, Aronson JK, Bankhead C, Mahtani KR, Nunan D. Confounding by indication. *Cat Biases* 2018. Available at: https://catalogofbias.org/biases/confounding-by-indication/. Accessed January 24, 2022.

6. Salas M, Hofman A, Stricker BHC. Confounding by indication: An example of variation in the use of epidemiologic terminology. *Am J Epidemiol* 1999; **149**: 981–983. doi:10.1093/oxfordjournals.aje.a009758.

7. Bousquet C, Henegar C, Louët AL Le, Degoulet P, Jaulent MC. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach. *Int J Med Inform* 2005; **74**: 563–571. doi:10.1016/j.ijmedinf.2005.04.006.

8. Pearson RK, Hauben M, Goldsmith DI, *et al.* Influence of the MedDRA® hierarchy on pharmacovigilance data mining results. *Int J Med Inform* 2009; **78**: 97–103. doi:10.1016/j.ijmedinf.2009.01.001.

9. Hill R, Hopstadius J, Lerch M, Noren GN. An attempt to expedite signal detection by grouping related adverse reaction terms. *Drug Saf* 2012; **35**: 1194–1195.

10. Bousquet C, Lagier G, Lillo-Le Louët A, Le Beller C, Venot A, Jaulent M-C. Appraisal of the MedDRA Conceptual Structure for Describing and Grouping Adverse Drug Reactions. *Drug Saf* 2005; **28**: 19–34.

11. Géniaux H, Assaf D, Miremont-Salamé G, *et al.* Performance of the standardised MedDRA® queries for case retrieval in the French spontaneous reporting database. *Drug Saf* 2014; **37**: 537–542. doi:10.1007/s40264-014-0187-2.

12. Wisniewski AFZ, Bate A, Bousquet C, *et al.* Good Signal Detection Practices: Evidence from IMI PROTECT. *Drug Saf* 2016; **39**: 469–490. doi:10.1007/s40264-016-0405-1.

13. 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* 2014; **37**: 557–567. doi:10.1007/s40264-014-0189-0.

14. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance. *Drug Saf* 2014; **37**: 655–659. doi:10.1007/s40264-014-0198-z.

15. Harpaz R, DuMouchel W, Shah NH. Comment on: "Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance." *Drug Saf* 2015; **38**: 113–114. doi:10.1007/s40264-014-0245-9.

16. Hauben M, Aronson JK, Ferner RE. Evidence of Misclassification of Drug–Event

Associations Classified as Gold Standard 'Negative Controls' by the Observational Medical Outcomes Partnership (OMOP). *Drug Saf* 2016; **39**: 421–432. doi:10.1007/S40264-016-0392-2/TABLES/7.

17. 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; **12**: 517–521. doi:10.1002/pds.885.

18. 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**: 539–46. doi:10.1038/clpt.2013.24.Performance.

19. 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**: 294–5. doi:10.1002/pds.

20. Strandell J, Caster O, Bate A, Norén N, Ralph Edwards I. *Reporting Patterns Indicative of Adverse Drug Interactions A Systematic Evaluation in VigiBase.*, 2011.

21. Hochberg AM, Hauben M, Pearson RK, *et al.* An evaluation of three signal-detection algorithms using a highly inclusive reference event database. *Drug Saf* 2009; **32**: 509–525. doi:10.2165/00002018-200932060-00007.

22. 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**. doi:10.1007/s40264-013-0097-8.

23. 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**: 561–575. doi:10.1007/s40264-016-0409-x.

24. Arnaud M, Bégaud B, Thiessard F, *et al.* An Automated System Combining Safety Signal Detection and Prioritization from Healthcare Databases: A Pilot Study. *Drug Saf* 2018; **41**: 377–387. doi:10.1007/s40264-017-0618-y.

25. Seabroke S, Candore G, Juhlin K, *et al.* Performance of Stratified and Subgrouped Disproportionality Analyses in Spontaneous Databases. *Drug Saf* 2016; **39**: 355–364. doi:10.1007/s40264-015-0388-3.

26. 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**: 13–23. doi:10.1007/s40264-012-0002-x.

27. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* 2010; **11**: S7. doi:10.1186/1471-2105-11-S9-S7.

28. Iyer S V., Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. *J Am Med Informatics Assoc* 2014; **21**: 353–362. doi:10.1136/amiajnl-2013-001612.

29. 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**. doi:10.1038/sdata.2016.26.

30. Kontsioti E, Maskell S, Dutta B, Pirmohamed M. A reference set of clinically relevant adverse drug-drug interactions. *Sci Data* 2022.

31. Harpaz R, Odgers D, Gaskin G, *et al.* A time-indexed reference standard of adverse drug reactions. *Sci Data* 2014; **1**: 140043.

32. Norén GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug–drug interaction surveillance. *Stat Med* 2008; **27**: 3057–3070. doi:10.1002/sim.

33. Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug-drug interactions in a spontaneous reports database. *Br J Clin Pharmacol* 2007; **64**: 489–495. doi:10.1111/j.1365-2125.2007.02900.x.

34. Norén GN, Caster O, Juhlin K, Lindquist M. Authors' Reply to Harpaz et al. Comment on: "Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance." *Drug Saf* 2015; **38**: 115–116. doi:10.1007/s40264-014-0246-8.

35. 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.

36. Maciejewski M, Lounkine E, Whitebread S, Farmer P, Shoichet BK, Urban L. The Powers and Perils of Post-Marketing Data Analysis: Quantification and Mitigation of Biases in the FDA Adverse Event Reporting System. doi:10.1101/068692.

37. 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**: 371–388. doi:10.1007/s40264-013-0053-7.

38. Hult S, Sartori D, Bergvall T, *et al.* A Feasibility Study of Drug–Drug Interaction Signal Detection in Regular Pharmacovigilance. *Drug Saf* 2020. doi:10.1007/s40264-020-00939-y.

Tables

Table 1: Medical concepts in the ***MC Reference Set***.

Name		
Acute kidney injury	Drug-induced liver injury	Myopathy
Acute psychosis	Hyperglycaemia	Priapism
Angioedema	Hypertension	Rhabdomyolysis
Arrhythmia	Hypoglycaemia	Tachycardia
Bradycardia	Hyponatraemia	Thrombocytopenia
Cardiac failure	Hypothyroidism	Torsade de pointes
Drug withdrawal syndrome	Lactic acidosis	

Table 2: Categories and descriptions of design criteria for reference sets that could affect performance evaluation of SDAs for DDI surveillance. The categories marked with an asterisk (*) contain design criteria that were not applicable to the ***MC Reference Set***.

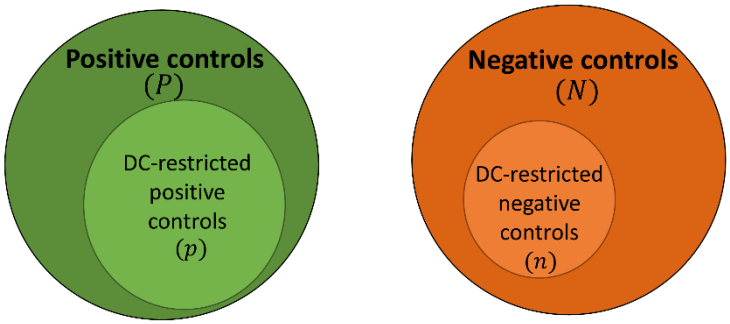
Category	Design Criterion (DC)	Description
Evidence level	*BNF - Study*	Interactions where the information is based on formal study including those for other drugs with same mechanism, e.g. known inducers, inhibitors, or substrates of cytochrome P450 isoenzymes or P-glycoprotein.
	BNF - Theoretical	Interactions 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.
	BNF – Anecdotal	Interactions based on either a single case report or a limited number of case reports.
	Micromedex – Established	Controlled studies have clearly established the existence of the interaction.
	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.
	Micromedex – Probable	Documentation strongly suggests the interactions exists, but well-controlled studies are lacking.
Event seriousness*	*EMA Important Medical Event (IME) Terms*	Any untoward medical occurrence that at any dose: * results in death, * is life-threatening, * requires inpatient hospitalisation or prolongation of existing hospitalisation, * results in persistent or significant disability/incapacity, or * is a congenital anomaly/birth defect.

	EMA Designated Medical Event (DME) Terms	Medical conditions that are inherently serious and often medicine-related (e.g., Stevens-Johnson syndrome). This list does not address product specific issues or medical conditions with high prevalence in the general population.
Event frequency*	*Common PTs*	PT prevalence ≥ 90th percentile of prevalence of PTs reported in FAERS
	Rare PTs	PT prevalence ≤ 10th percentile of prevalence of PTs reported in FAERS
Potential confounding by indication	*AE is an indication - True*	The AE is also an indication for at least one of the two drugs from the drug-drug-event triplet under consideration
	AE is an indication - False	The AE is not an indication for either of the drugs from the drug-drug-event triplet under consideration
Potential confounding by concomitant medication	*Shared indications – False*	Drug pairs that share at least one indication are excluded
	Shared indications – True	Only drug pairs that share at least one indication are considered

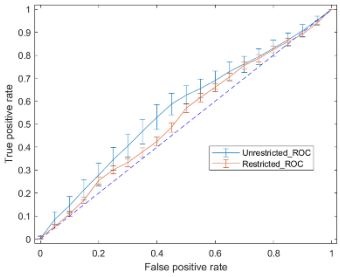
Table 3: Statistics related to the performance evaluation of three SDAs for DDIs using matched controls from the ***PT Reference Set*** and ***MC Reference Set***.

SDA	***PT Reference Set*** **AUC (95% CI)**	***MC Reference Set*** **AUC (95% CI)**
Omega	0.6011 (0.5704, 0.6317)	0.5406 (0.5150, 0.5662)
delta_add	0.4645 (0.4408, 0.4882)	0.4956 (0.4721, 0.5191)
IntSS	0.5374 (0.5100, 0.5648)	0.4885 (0.4654, 0.5117)

Figures



(a)



Restricted reference set (1:1) sizes : *100; 200; 1000; 2000; P* or *N (max)*

1. Random sampling from p and n → *Restricted_ROC*
2. Random sampling from P and N → *Unrestricted_ROC*

1,000 simulations

$$\boldsymbol{AUC_{diff} = AUC_{Restricted\_ROC} - AUC_{Unrestricted\_ROC}}$$

$$|AUC_{diff}| \sim N\left(|\mu_{AUC_{Restricted\_ROC}} - \mu_{AUC_{Unrestricted\_ROC}}|, \sqrt{\sigma_{AUC_{Restricted\_ROC}}^{2} + \sigma_{AUC_{Unrstricted\_ROC}}^{2}}\right) \quad (1)$$

$$P(|AUC_{diff}| > 0) = 1 - P(|AUC_{diff}| = 0) = 1 - F_{AUC_{diff}}(0) \quad (2)$$

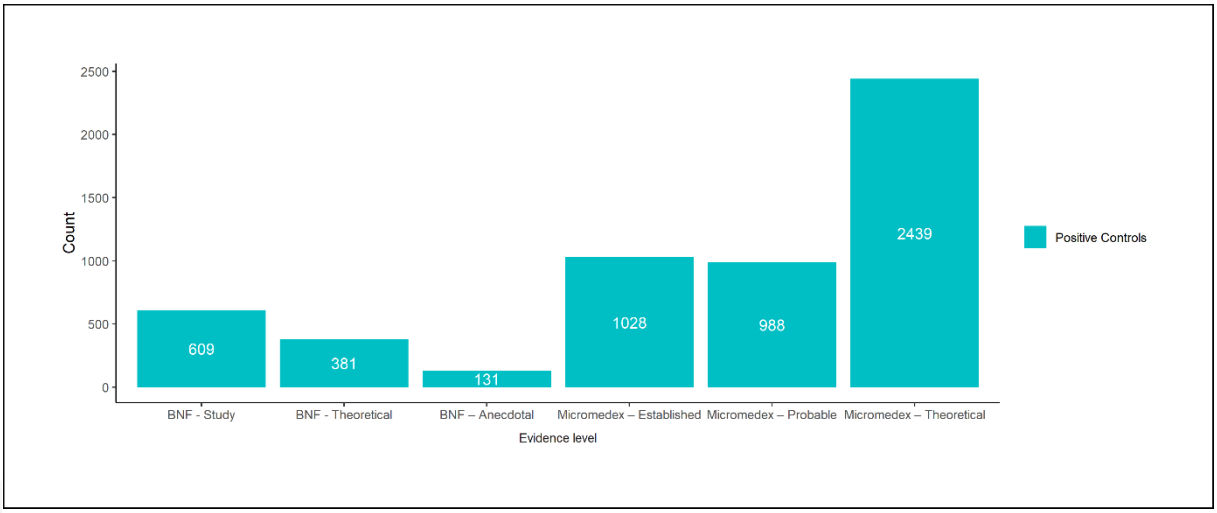
For fixed sensitivity values (0.60; 0.75; and 0.90): $\boldsymbol{PPV_{diff} = PPV_{Restricted\_ROC} - PPV_{Unrestricted\_ROC}}$

* μ: mean; σ : standard deviation;
$F_{AUC_{diff}}$: normal cumulative distribution function (CDF) of AUC_{diff}

(b)

Figure 1: (a) Initial positive and negative control sets (P and N) and their respective restricted subsets (*DC-restricted*, p and n) when applying a design criterion; (b) Simulation workflow for the of differences in AUC (AUC_{diff}) and PPV (PPV_{diff}) when considering the specified design criterion.

(a)



(b)

Figure 2: (a) Number of positive and negative controls from the ***PT Reference Set*** for each of the different design criteria when the restricted subsets contained both control types. The maximum number of controls considered from each type to form simulated reference sets ($\boldsymbol{N_{max}}$) is denoted with white color in the respective bar; (b) Number of ***PT Reference Set*** positive controls for the *Evidence level* design criteria, where restriction could not be applied to negative controls.

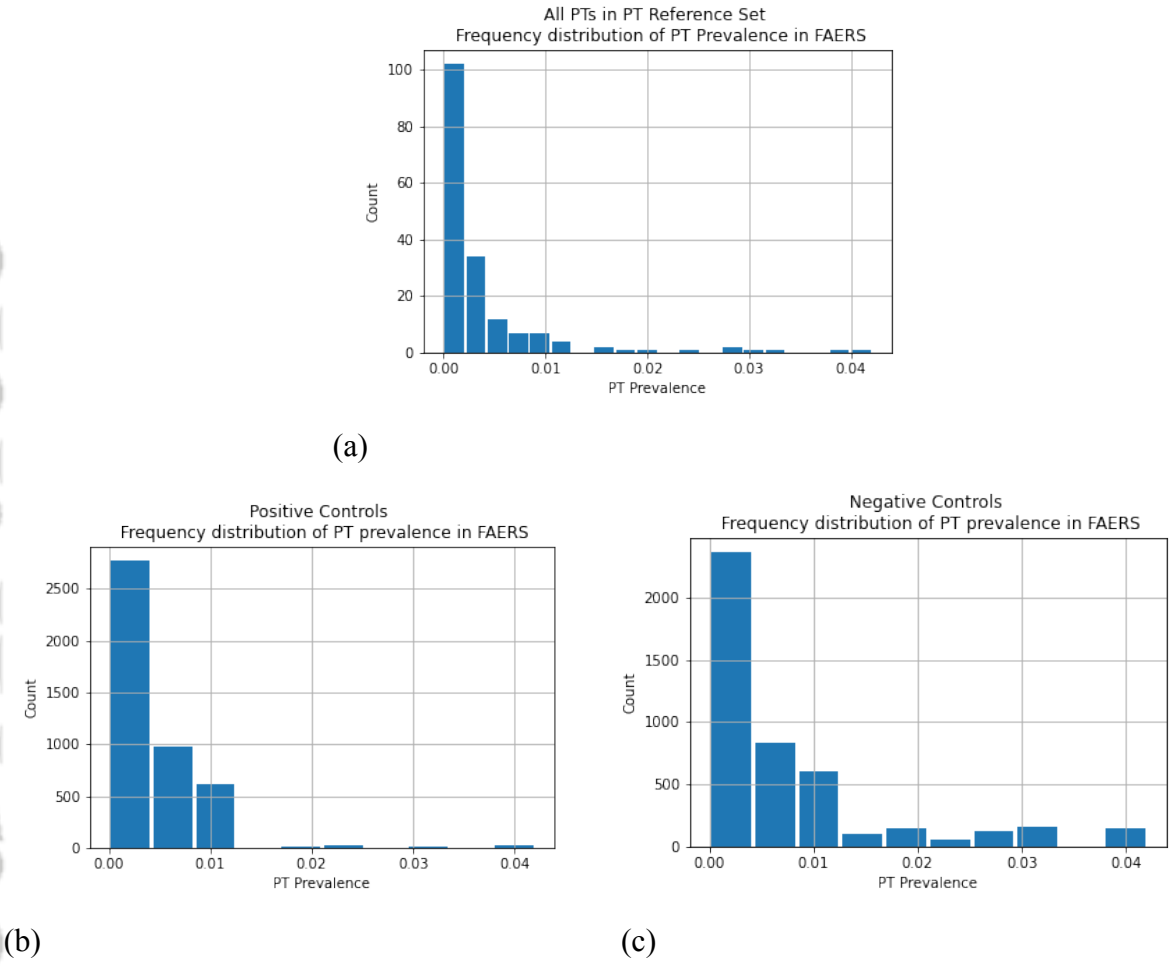


Figure 3: Frequency distribution of PT prevalence in FAERS for: (a) the set of unique PTs in the ***PT Reference Set***; (b) PTs contained in the ***PT Reference Set*** positive controls; and (c) PTs contained in the ***PT Reference Set*** negative controls.

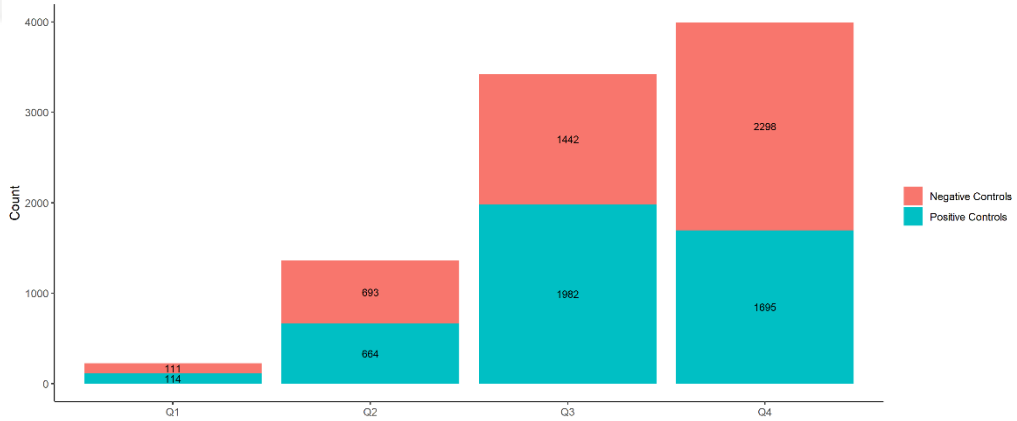
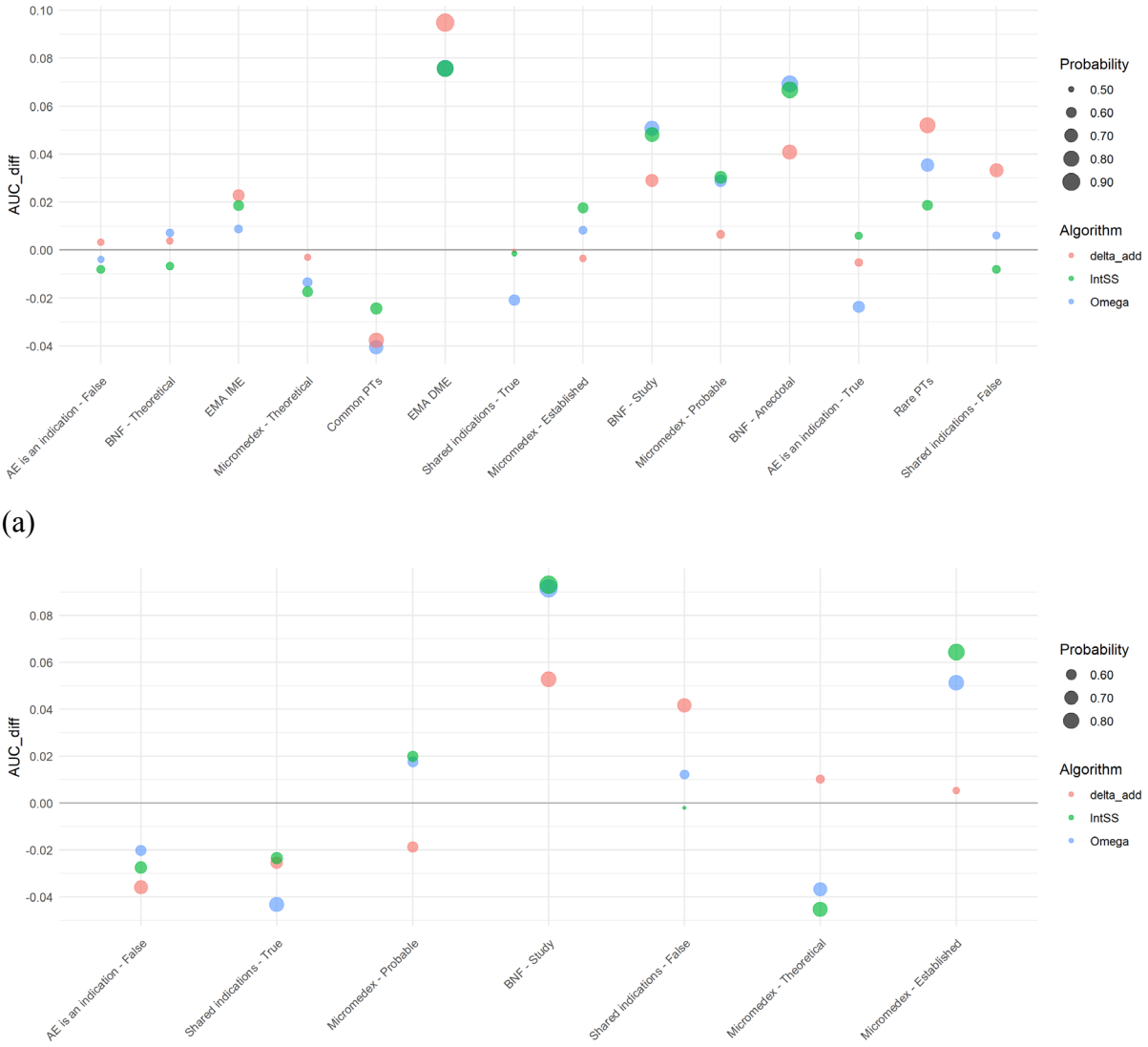


Figure 4: Number of positive and negative controls for groups Q1 to Q4 that were formed using PT prevalence quartile binning, with Q1 containing the controls with the lowest prevalence and Q4 the highest one.



(a)

(b)

Figure 5: AUC_{diff} for a fixed restricted reference set size of 100 for: (a) the ***PT Reference Set***; (b) the ***MC Reference Set***. Design criteria are ordered by increasing range of AUC_{diff} values among the three SDAs. The dot size represents the probability of the estimated score,

AUC_{diff}, being non-zero.

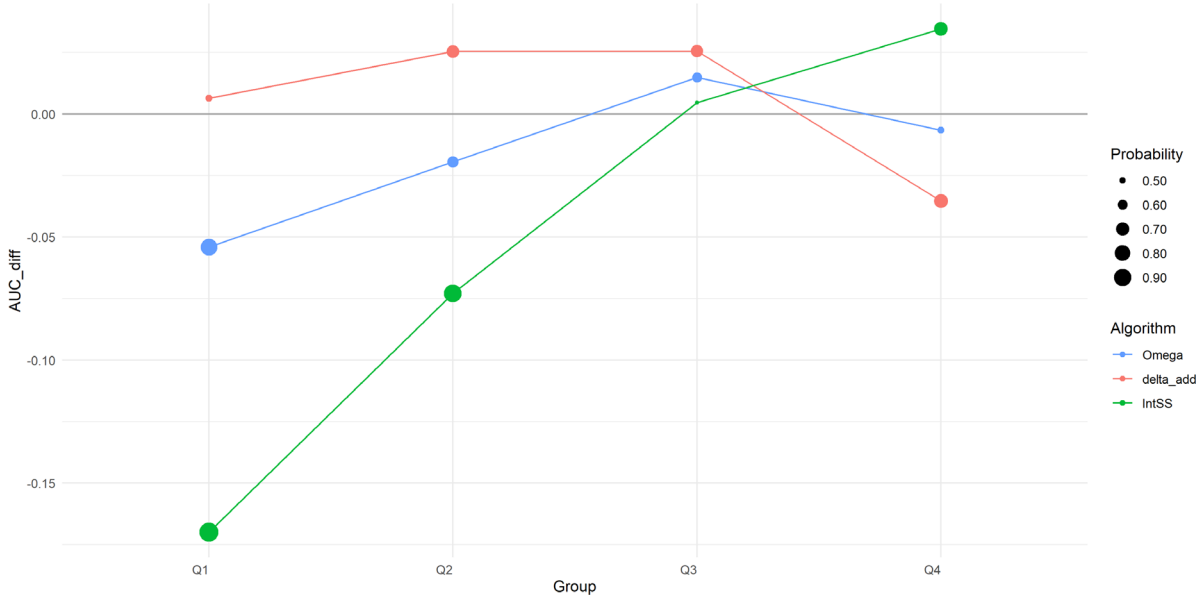


Figure 6: AUC_{diff} values for *Groups Q1* to *Q4* relevant to PT prevalence. The dot size represents the probability of the estimated score, AUC_{diff}, being non-zero.