Drug Saf
DOI 10.1007/s40264-015-0388-3

ORIGINAL RESEARCH ARTICLE

Performance of Stratified and Subgrouped Disproportionality Analyses in Spontaneous Databases

Suzie Seabroke[1] · Gianmario Candore[2] · Kristina Juhlin[3] · Naashika Quarcoo[4] · Antoni Wisniewski[5] · Ramin Arani[5] · Jeffery Painter[4] · Philip Tregunno[1] · G. Niklas Norén[3] · Jim Slattery[2]

Abstract

Introduction Disproportionality analyses are used in many organisations to identify adverse drug reactions (ADRs) from spontaneous report data. Reporting patterns vary over time, with patient demographics, and between different geographical regions, and therefore subgroup analyses or adjustment by stratification may be beneficial.

Objective The objective of this study was to evaluate the performance of subgroup and stratified disproportionality analyses for a number of key covariates within spontaneous report databases of differing sizes and characteristics.

Methods Using a reference set of established ADRs, signal detection performance (sensitivity and precision) was compared for stratified, subgroup and crude (unadjusted) analyses within five spontaneous report databases (two company, one national and two international databases). Analyses were repeated for a range of covariates: age, sex, country/region of origin, calendar time period, event seriousness, vaccine/non-vaccine, reporter qualification and report source.

Results Subgroup analyses consistently performed better than stratified analyses in all databases. Subgroup analyses also showed benefits in both sensitivity and precision over crude analyses for the larger international databases, whilst for the smaller databases a gain in precision tended to result in some loss of sensitivity. Additionally, stratified analyses did not increase sensitivity or precision beyond that associated with analytical artefacts of the analysis. The most promising subgroup covariates were age and region/country of origin, although this varied between databases.

Conclusions Subgroup analyses perform better than stratified analyses and should be considered over the latter in routine first-pass signal detection. Subgroup analyses are also clearly beneficial over crude analyses for larger databases, but further validation is required for smaller databases.

Key Points

Subgroup analyses perform better than stratified analyses for routine first-pass signal detection.

There are clear benefits of subgroup analyses over crude analyses for large international databases, whilst smaller databases may need to consider a trade-off in performance characteristics.

✉ Suzie Seabroke
suzie.seabroke@mhra.gsi.gov.uk

1 UK Medicines and Healthcare Products Regulatory Agency (MHRA), 151 Buckingham Palace Road, London SW1W 9SZ, UK

2 European Medicines Agency, London, UK

3 Uppsala Monitoring Centre, Uppsala, Sweden

4 GlaxoSmithKline, London, UK

5 AstraZeneca, Alderley Park, UK

1 Introduction

Spontaneously reported adverse drug reaction (ADR) data have been an important source of drug safety information for over 50 years [1]. The basis for the data is the identification and reporting of suspected adverse reactions to drugs/vaccines by healthcare professionals and patients

Published online: 09 January 2016

which are then collected and analysed by manufacturers, regulatory agencies and independent drug safety monitoring organisations. Spontaneous reports reflect suspicions of ADRs and are not necessarily indicative of a causal relationship between a drug and an event. The challenge is distinguishing ADRs from spurious associations. With increasing volumes of reports received every year, medical assessment of individual reports becomes impossible for many organisations. Thus, the use of statistical disproportionality methods to aid signal detection from spontaneous report data has become well-established [2–6].

Spontaneous report databases cover a range of products aimed at diverse medical conditions/indications and are used across a broad range of patient populations. For example, vaccines are given to healthy subjects, especially children who are likely to have fewer underlying medical conditions and consequently a differing rate of reported background adverse events than the main population of patients that use other medicines. Many statistical signal detection methods disregard this diversity and give equal weight to information from all products and all patients when computing the expected number of reports for a particular drug–event pair. However, ignoring the diversity within the dataset may result in signals either being masked or false associations being flagged as potential signals through either confounding or effect modification [7]. Stratification is generally used in epidemiology to reduce confounding by dividing data into groups or strata that have the same value of the confounding factor. Stratum-specific estimates are then pooled from each strata to provide an overall estimate. Stratified analyses assume that there is no variation in risk across strata. Where there is effect modification, i.e. variation in risk exists between strata, then analysis of the data within subgroups is conducted to preserve and highlight this variation. Both of these approaches may also have advantages in statistical signal detection.

There is increasing interest in whether stratified or subgroup analyses could provide additional benefit to the established statistical signal detection methods, with some organisations already routinely using these analyses [8–13]. A few studies have investigated the impact of stratification [7, 11–16] and subgroup analyses [11, 17, 18] on signal detection algorithms. The studies were conducted in a range of different databases and focused on the impact of stratification and/or subgrouping on a few key covariates including age, sex, time period, country of origin, reporter qualification, type of report, vaccine/non-vaccine and therapeutic drug class. The findings were that stratified analyses generally highlighted fewer drug–event pairs, which may have some benefit in increasing the efficiency of signal detection algorithms, although this assumes that the signals no longer highlighted in the stratified analysis are indeed false positives in the crude analysis [14–16]. Some studies evaluated signal detection performance through measuring the ability to detect known ADRs using a reference standard of known ADRs (e.g. ADRs listed in product information or overviews of published case reports). Modest improvements in performance were observed for stratified analyses compared with crude analyses, with some covariates having a greater impact than others [7, 13]. The study by Hopstadius et al. [7] highlighted the risk of over-stratification if too many variables are adjusted for simultaneously generating small strata. This study observed a loss of sensitivity in the presence of small strata. Studies investigating subgrouping have also found some benefits, but particularly for the vaccine/non-vaccine subgroups they have found that signals could also be missed [16–18]. A further study by Hopstadius and Norén [11] found that a large number of potential ADRs could be uncovered by stratified or subgroup analyses or a combination of these methods compared to crude analyses, and that subgroup analyses uncovered more drug–event pairs than stratified analyses. However, this study did not evaluate signal detection performance against a reference standard.

The evidence from previous studies has suggested some benefits of stratified and subgroup analyses but often the analyses included only a few key covariates or study products and were conducted in single databases. It is not clear how generalisable these results are to other spontaneous report datasets of different sizes and characteristics. Additionally, to our knowledge, a head-to-head comparison of stratified and subgroup analyses against a reference standard has not been conducted. This study therefore aimed to investigate the impact of stratified and subgroup analyses for routine first-pass signal detection within several spontaneous report datasets of varying size and characteristics using a wide range of key covariates with signal detection performance measured against a reference standard. The overall objective of the study was to provide some conclusions on whether these analyses are truly beneficial to routine statistical signal detection and, if so, in what circumstances they should be used.

2 Methods

2.1 Data Sources

A total of five different spontaneous report databases were analysed using data up to 31 December 2011: VigiBase® (WHO global individual case safety reports database, 7.0 million reports), EudraVigilance (European Medicines Agency [EMA] database of reports from pharmaceutical companies and European regulatory agencies, 2.4 million

Table 1 Characteristics of participating databases

Organisation (abbreviation)	Affiliation	Database name	Number of spontaneous reports (millions)[a]	Coverage	Number of products included in study
Uppsala Monitoring Centre (UMC)	Drug safety monitoring agency	VigiBase®	7.0	Global	220
European Medicines Agency (EMA)	Competent authority	EudraVigilance	2.4	Pan-European[b]	220
Medicines and Healthcare products Regulatory Agency (MHRA)	Competent authority	Sentinel	0.6	UK	207
GlaxoSmithKline (GSK)	Pharmaceutical company	OCEANS	1.4	Global	21
AstraZeneca (AZ)	Pharmaceutical company	Sapphire	0.5	Global	11

[a] Data lock point for total report counts: VigiBase®, Sentinel, Sapphire—30 June 2010; EudraVigilance—2 December 2010; OCEANS—20 May 2011

[b] Also includes serious unexpected reports from the rest of the world

reports), Medicines and Healthcare products Regulatory Agency (MHRA) Sentinel database (UK regulatory agency database, 0.6 million reports), OCEANS (GlaxoSmithKline's worldwide safety database, 1.4 million reports) and Sapphire (AstraZeneca's worldwide safety database, 0.5 million reports). The main features of the participating databases are summarised in Table 1.

The analysis was carried out separately in each different database but using a common methodology to ensure the results were directly comparable.

2.2 Covariates

Spontaneous report data contain information on many variables other than the drug–event of interest, which could theoretically be used as stratification/subgroup covariates. Not all data fields in spontaneous reports are always completed, however, leading to missing information which may affect the usefulness of these variables for stratified or subgroup analysis. Furthermore the study by Hopstadius et al. [7] highlighted the need to avoid over-stratification, which may result in reduced sensitivity. This study investigated the impact of stratification/subgrouping of a number of key covariates (Table 2). A strategy for avoiding small strata was used for country of origin that grouped countries with $\leq$100 reports by region/ethnicity. Sensitivity analyses were also conducted to either include or exclude data with missing covariate information (age, sex). A number of combinations of variables were also analysed for covariates with the most promising initial results to investigate if performance could be further enhanced. All databases contributed to the analyses as far as was possible, e.g. MHRA could not contribute to the analyses on region/country of origin as only UK data are available and AstraZeneca and the EMA could not contribute to the vaccine subgroup analysis.

2.3 Statistical Analysis

It was envisaged that the effect of stratification or subgrouping would be similar between different disproportionality methods [19]. For comparability, the analysis was carried out across all databases using the reporting odds ratio (ROR) [20] with statistical signal criteria as follows: lower bound 95 % confidence interval $\geq$1, $n \geq 3$. When these criteria are met it is termed a signal of disproportionate reporting (SDR). To ensure the results were generalisable to most signal detection methods, analyses were also replicated using Bayesian methods: the information component (IC) [21] with a statistical signal criterion of IC025 (lower bound of two-sided 95 % credibility interval) >0 in four databases; and MGPS (multi-item gamma poisson shrinkage) [22] with a signal criteria of empirical Bayes geometric mean (EBGM) $\geq$2.5, EB05 $\geq$1.8 and $n \geq 3$ in one database. The thresholds were chosen as those in current use within pharmacovigilance departments.

Stratified analyses were conducted using a Mantel–Haenszel estimate for the odds ratio for the ROR [23] and a Mantel–Haenszel type of adjustment for the IC and EBGM which stratifies the computation of expected counts and then replaces the overall expected count with a sum over stratum-specific expected counts [21]. Subgroup analyses calculated disproportionality measures within each individual stratum separately and a statistical signal was counted if the score from any of the strata met the signal criterion. For covariates such as age and sex that have the potential for missing data, the subgroup analyses did not include a separate category for missing data but excluded these reports from the analysis. Results from the stratified and subgroup analyses were compared to the crude unadjusted results.

Additional analyses were also conducted. One was to investigate whether the inclusion of missing data for age

Table 2 Covariates included in the study

Covariate	Strata
Age	0–23 months, 2–11, 12–17, 18–35, 36–64, 65–74, 75+ years, unknown
Sex	Male, female, unknown
Time period	5-yearly
Vaccines/drugs	Vaccines, non-vaccines
Event seriousness	Serious, non-serious[a]
Reporter qualification	Consumer only, healthcare professional only, mixed
Report source	Spontaneous[b], solicited/legal cases
Country of origin	Individual country of origin[c]
Region of origin	North America, Europe, Japan, rest of Asia, rest of the world

[a] Based on European Medicines Agency Intensively Monitored Event (IME) list (available at http://eudravigilance.ema.europa.eu/human/textforIME.asp)

[b] Definition of spontaneous varies across the different databases. Individual definitions are available in Candore et al. [19]

[c] To avoid small strata, countries were grouped by region/ethnicity if $n < 100$: Central and Southern Africa, North Africa and Middle East, Central Asia, East Asia, Southeast Asia, Australasia, Caribbean, South America, Central Europe, Eastern Europe, Scandinavia and Iceland, Southern Europe, UK and Ireland, North America and Canada

and sex as a separate stratum had an impact on the results of the stratified analyses. Another was to determine whether a modified signal criterion for subgroup analyses, where the threshold for the minimum number of reports was based on all reports for the drug–event combination rather than the reports within each stratum individually, might also have an effect. These additional analyses were not conducted in all databases and not all disproportionality methods were included.

A final sensitivity analysis investigated whether any effect observed by stratifying or subgrouping could be explained by analytical artefacts associated with the stratification or subgroup process. Data were re-analysed through a permutation analysis using randomly split strata of equal size to a real variable of interest (age, country of origin and reporter) and compared to the results for the real variable [11]. This analysis was performed in the two largest databases: EudraVigilance and VigiBase®.

2.4 Performance Evaluation

The overall aim of statistical signal detection methods is to detect ADRs as quickly as possible whilst minimising false positives. The outcome measures used in this study were sensitivity (the proportion of known ADRs that are correctly highlighted) and precision (the proportion of SDRs that correspond to a known ADR). Whilst timing is clearly also an important factor in signal detection, this initial analysis concentrated on whether an ADR was ever highlighted and the proportion of false positives generated. The sensitivity and precision were calculated using data to the end of December 2011.

The reference set and product list used in this study are the same as those used in a previous PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium) study comparing the performance of crude (i.e. unstratified/not subgrouped) statistical signal detection methods within and across different databases [19].

The reference standard for determining if an SDR was an ADR or not was based on section 4.8 of the summary of product characteristics (SPC) and company reference safety information. All SDRs that did not correspond to a term in the reference dataset were defined as false positives.

Due to the resource implications to obtain a reference dataset for all known medicinal products it was not feasible to conduct this study for all products with reports in the spontaneous databases. A list of 220 study products was therefore selected to include a broad range of products from different therapeutic areas. The list was initially based on the study products used previously in a proportional reporting ratio validation study carried out by the EMA [24] with additional products included to ensure a representative proportion of drugs used in different age groups and to ensure sufficient numbers of products for analysis in each company database. The choice of study products was determined without reference to the product information that would become the reference dataset. Not all partner databases were able to include all study products, however, particularly the company databases that only contain reports for their own products.

3 Results

The effect of using either stratification or subgrouping in disproportionality analyses is shown in Figs. 1 and 2 split by database. Figure 1 shows the results for the ROR

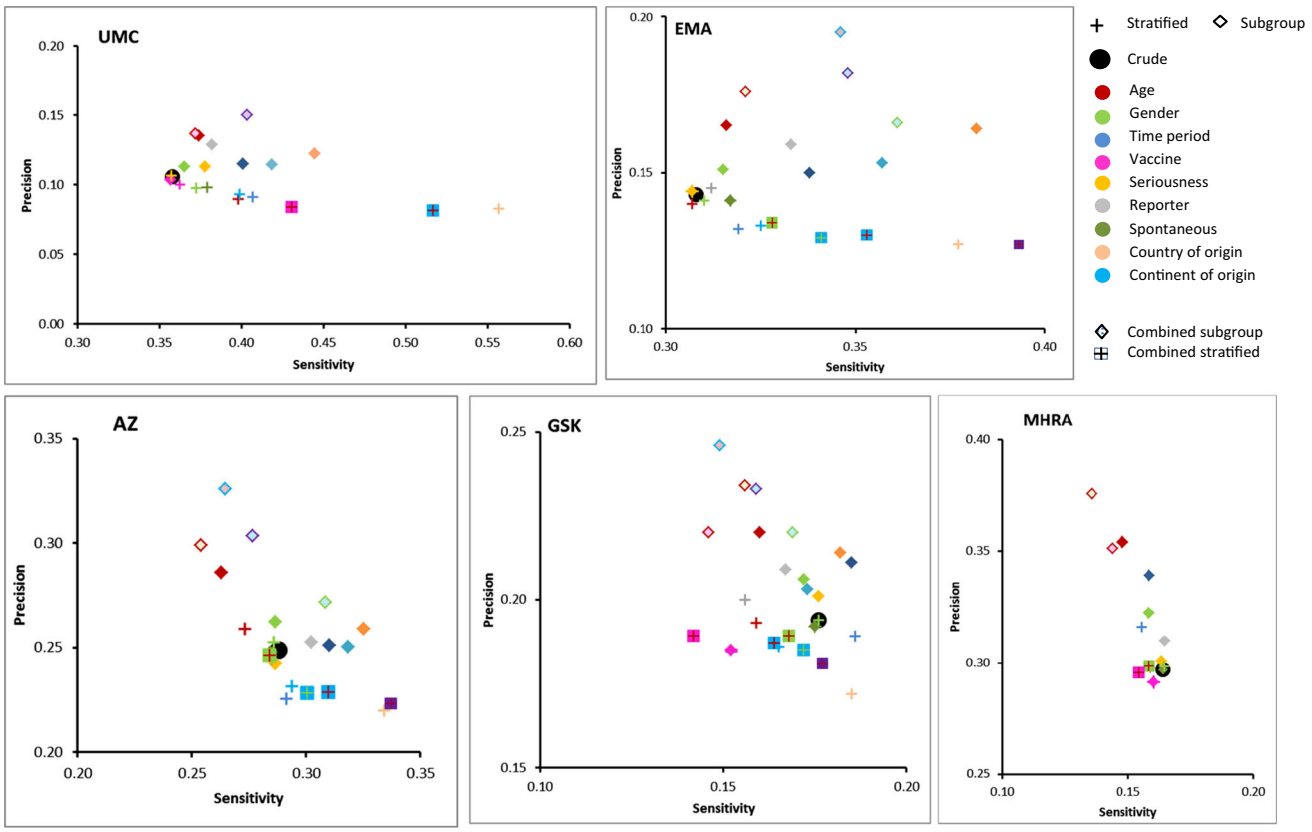

Fig. 1 Sensitivity and precision for crude, stratified and subgroup reporting odds ratio analyses. *AZ* AstraZeneca, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *MHRA* Medicines and Healthcare products Regulatory Agency, *UMC* Uppsala Monitoring Centre

analyses and Fig. 2 shows the results for the Bayesian methods (IC and MGPS). Results are presented as the sensitivity and precision for the crude, stratified and subgroup analyses.

Overall, subgroup analyses tended to perform (in terms of both sensitivity and precision) better than stratified analyses. This was particularly evident for the two largest databases (EudraVigilance and VigiBase®) where subgroup analyses performed better than stratified analyses for all variables. Some stratified analyses showed modest improvements in either sensitivity or precision from the crude analysis but not both, whereas the corresponding subgroup analysis was seen to achieve a higher sensitivity or precision or even both in some databases. This effect was seen for both the ROR and Bayesian disproportionality methods. Within the larger international datasets, consistent benefits in both precision and sensitivity for subgroup analyses over crude analyses were observed for the two disproportionality methods/thresholds but the Bayesian subgroup analysis tended to have lower improvements in sensitivity than the ROR subgroup analysis. For the smaller databases a gain in precision tended to result in some loss of sensitivity, particularly for the stricter Bayesian methods/thresholds and for the regulatory dataset in the UK with reports from only one country.

The results of the permutation analysis are presented in Fig. 3 as the absolute difference in sensitivity and precision from the crude analyses and showed that the effect of stratification by randomly split strata matched almost exactly that of stratification by the real covariates. The same analysis for the subgroup analyses, however, showed improvements, with the results for the real variables having a similar increase in precision to the permutation analysis but an improved sensitivity.

Subgrouping by age, country or region of origin or a combination of these variables showed the highest improvement in precision in all spontaneous report databases and also sensitivity in the larger databases (Figs. 1, 2). Subgrouping by sex, reporter type and 5-yearly time-points showed a modest improvement in precision for all databases and some improved sensitivity for larger and international databases. Subgrouping by seriousness of the event had little effect on either sensitivity or precision in any database and the analysis excluding legal cases from the dataset also had little effect in all databases apart from VigiBase®. Subgrouping by vaccines/non-vaccines resulted in a decrease in both precision and sensitivity in all spontaneous report databases that contributed data. Further investigation into this finding carried out on the MHRA Sentinel database restricting the study products to vaccines

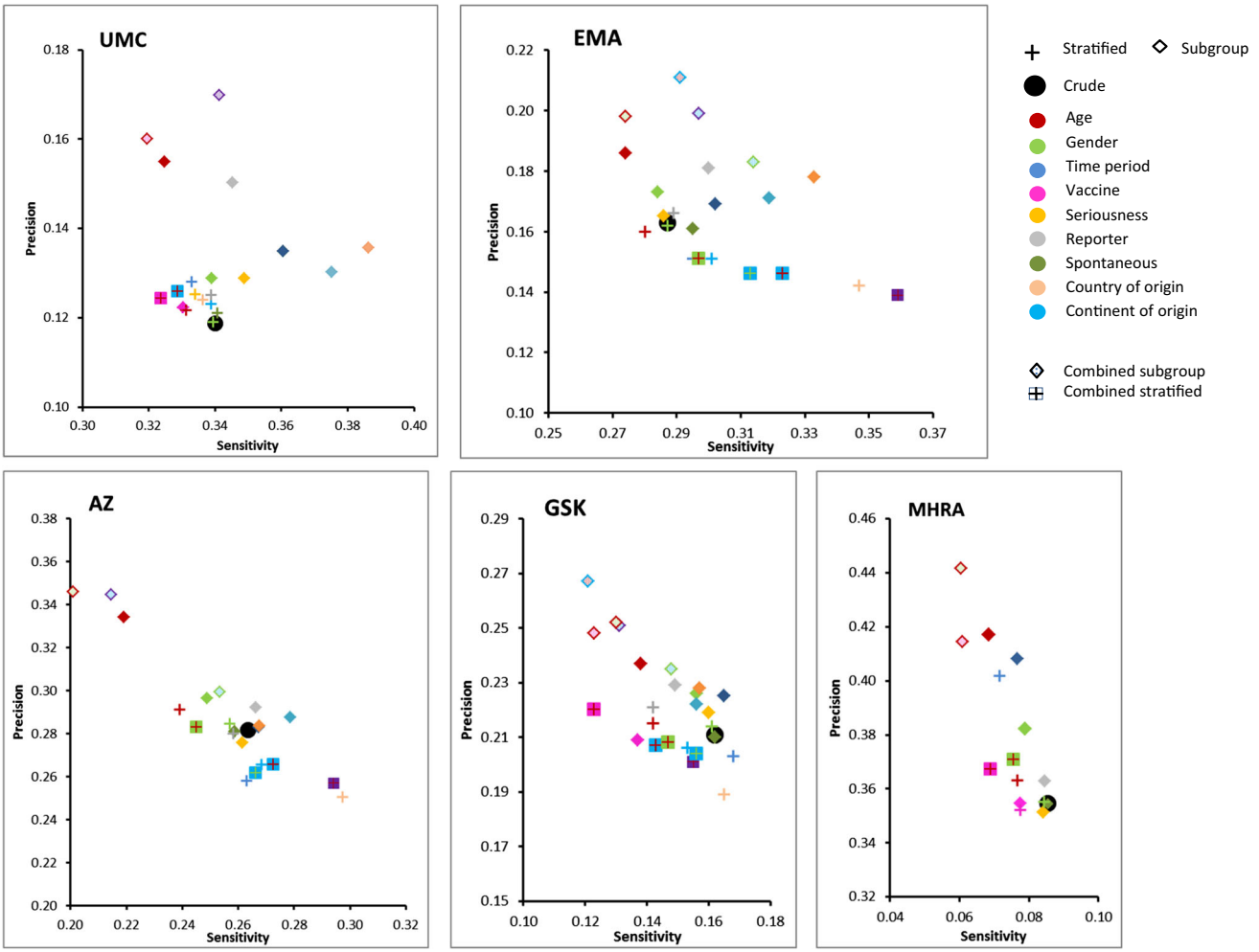

Fig. 2 Sensitivity and precision for crude, stratified and subgroup Bayesian analyses. *AZ* AstraZeneca, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *MHRA* Medicines and Healthcare products Regulatory Agency, *UMC* Uppsala Monitoring Centre

or non-vaccines only revealed that this result was almost exclusively driven by the vaccines subgroup rather than the non-vaccine subgroup (Fig. 4). A qualitative assessment of the events either missed or gained using either a vaccine-only subgroup or crude approach for vaccines also revealed that the crude analysis would detect many reactogenic-type events common to vaccines in addition to some more serious events such as Guillain–Barré syndrome. The vaccine subgroup suppressed these events and tended to highlight a more diverse range of events, e.g. cardiac events, laboratory test results, etc.

Including an additional category for missing data in the subgroup analyses for age and sex slightly increased the sensitivity in all databases but tended to also decrease precision when compared with the same analysis that excluded these data. In the databases with higher levels of missing data for these variables (≥20 %), the increase in sensitivity was greater than the decrease in precision (data not shown). Results for the additional subgroup analyses carried out that applied the minimum number of reports criteria to the whole drug–event combination rather than within each strata showed large increases in sensitivity in all databases but with some loss of precision in most databases for the ROR (Fig. 5). This analysis was replicated for the ROR increasing the minimum number of reports from three to five in three databases, with very similar results to the original analysis but with some reduction in sensitivity and very small increases in precision observed. However, the same analysis for the MGPS disproportionality method in one database showed little difference from the main subgroup analyses, suggesting that the number of reports has relatively little influence as a signalling criterion within this implementation of the method.

4 Discussion

Subgroup analyses consistently performed better (in terms of both sensitivity and precision) than stratified analyses for all of the covariates investigated in the two largest

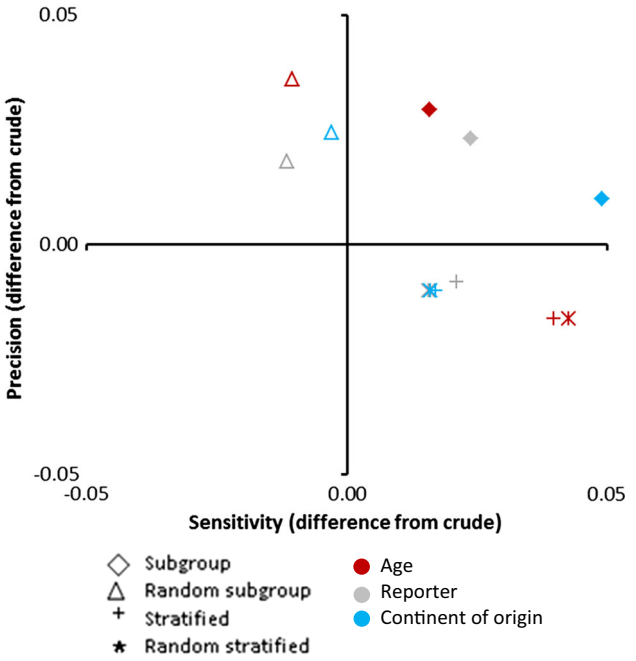

Fig. 3 Sensitivity and precision (absolute difference from crude) for stratified, subgroup and permutation reporting odds ratio analyses. Data are from the European Medicines Agency (continent of origin) and Uppsala Monitoring Centre (age, reporter) databases

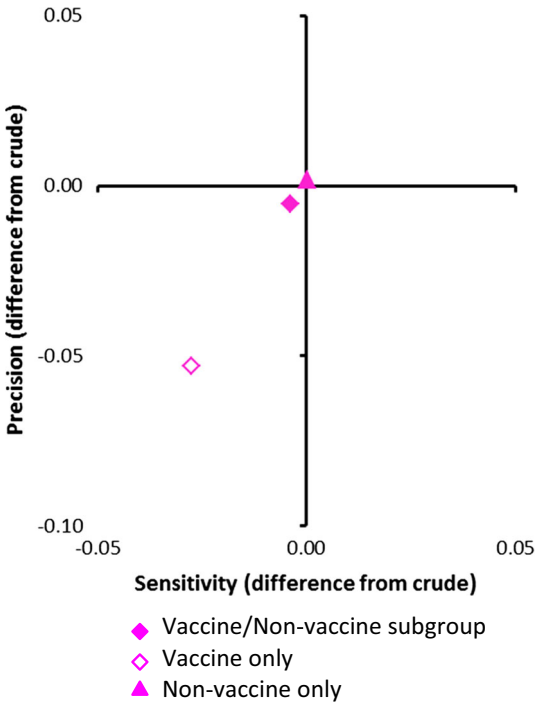

Fig. 4 Sensitivity and precision (absolute difference from crude) for vaccine/non-vaccine subgroup analyses for reporting odds ratio. Data are from the UK Medicines and Healthcare products Regulatory Agency database

databases (Figs. 1, 2). The pattern of performance did vary between databases, however, with a higher precision but lower sensitivity observed for the smaller databases for some covariates. This finding was seen for both the ROR and Bayesian disproportionality methods. Additionally, the results for the stratified analyses were shown to be consistent with artefacts from the analytical approach, as reflected by the similar results yielded in the permutation analysis, whereas the subgroup analyses did provide evidence of a benefit in sensitivity beyond random variation. This would indicate that whilst there may be confounding in the dataset for certain drugs and events and stratification may be beneficial in these examples, as observed by Woo et al. [14], adjusting for key variables across the whole dataset does not provide an overall benefit. This is consistent with the suggestion by Bate et al. [25] that routine use of stratification in signal detection cannot account for all confounding factors and may reduce the potential for early detection of signals. Other previous studies [7, 13] have observed modest improvements for stratified analyses consistent with the results from the stratified analyses in our study and therefore these findings from the other studies may also be artefacts from the stratification process rather than a true effect. The potential vulnerability of stratification to data quality issues is also highlighted by Hopstadius et al. [26].

In the two largest databases (VigiBase® and EudraVigilance), subgroup analyses improved both sensitivity and precision over crude analyses for most of the covariates, with age, region and country of origin or a combination of these providing the largest benefit. These findings suggest that a variation in the reporting rate across different subgroups exists in spontaneous data frequently enough that overall application of subgroup analyses to the entire dataset has benefits (at least for the larger databases).

The permutation analysis, however, indicates that only the increased sensitivity is a true beneficial effect of the subgroup analysis. The increase in precision from the crude analysis observed for the permutation/real analysis most likely correlates with the increased total number of reports required for an association to be detected in a specific subgroup. This effect is observed in the sensitivity analysis that applied the minimum number of reports criterion to the whole drug–event combination rather than within each stratum and tended to show large increases in sensitivity but a loss of precision. In the three smaller databases, an increase in precision tended to result in some loss of sensitivity, although an increase in both was observed for some variables (country of origin, time and reporter) in the two pharmaceutical industry databases. Although the

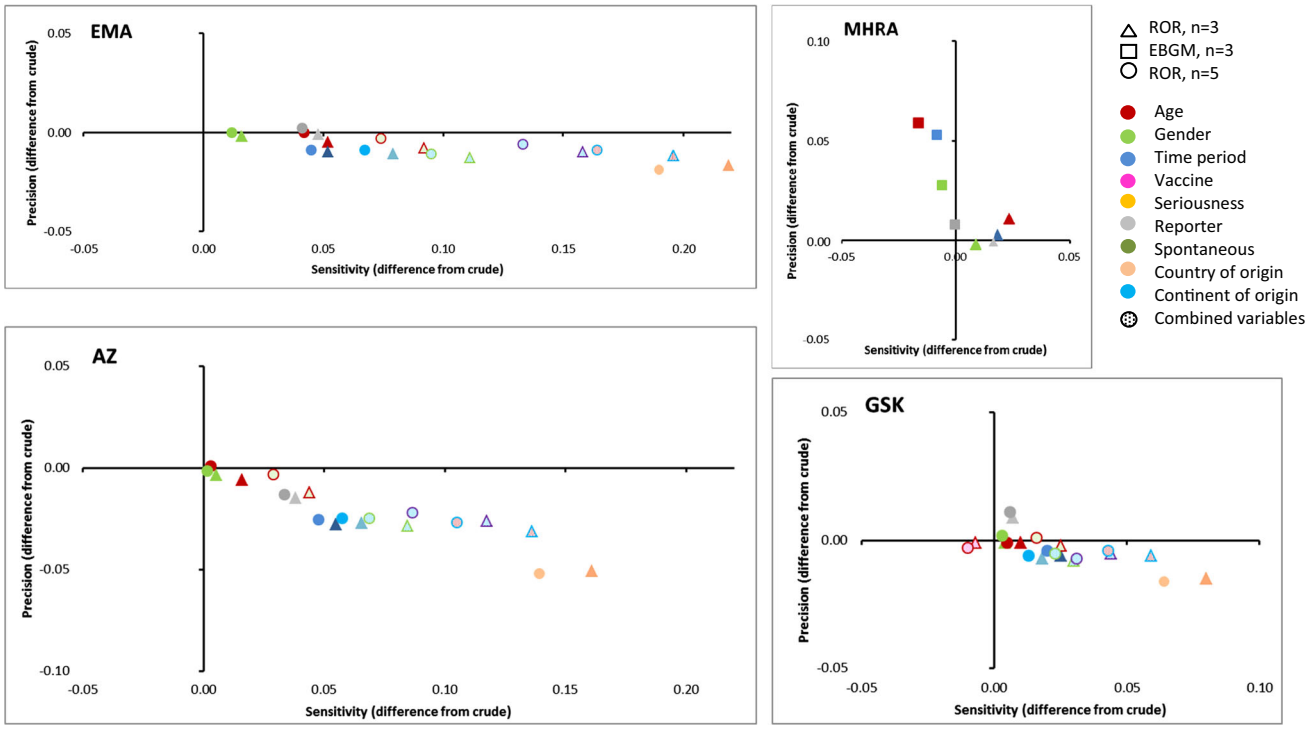

Fig. 5 Sensitivity and precision (absolute difference from crude) for subgroup ROR analyses where minimum count is applied across the whole drug–event combination. *AZ* AstraZeneca, *EBGM* empirical Bayes geometric mean, *EMA* European Medicines Agency, *GSK* GlaxoSmithKline, *MHRA* Medicines and Healthcare products Regulatory Agency, *ROR* reporting odds ratio

permutation analysis was only performed in the two largest databases, it seems likely that some or all of the increased precision observed in the smaller databases may be an artefact. If this was indeed the case then the benefit of subgroup analyses in smaller databases is questionable, with only a few covariates potentially having a beneficial effect on sensitivity. The choice of covariates to include in any signal detection subgroup analysis will likely need to vary between databases and organisations will need to take into account a potential trade-off between precision and sensitivity when deciding which variables to use in any subgroup analysis.

The vaccine/non-vaccine subgroup was the only covariate to show an overall decrease in both precision and sensitivity in all three databases that were able to contribute vaccine data. This result is not surprising as a vaccine subgroup will compare reports of common vaccine ADRs such as injection-site reactions with a background of other vaccines where these types of reactions are common, resulting in lower disproportionality. Such reactions are, however, those that are labelled in the product information and that are counted as true positives in this study. This has been observed in other studies [13, 16] and confirmed by the additional analyses conducted in the MHRA database that showed that the vaccine-only subgroup did indeed detect fewer reactogenic-type reactions and other reactions common to vaccines. Whilst it may be desirable to suppress commonly reported labelled vaccine reactions such as injection-site reactions within a routine signal detection system, consideration must be given to the more serious reactions such as Guillain–Barré syndrome that will also be suppressed. Crude and subgroup approaches used in parallel may be more appropriate for vaccines.

The results from the additional analyses conducted also highlight areas to be considered by organisations prior to implementing subgroup analyses into routine signal detection. The subgroup analyses for age and sex that included an additional category for missing data showed improvements in sensitivity with some loss of precision compared with the analysis that excluded the missing data. This loss in precision was reduced for databases with higher levels ($\geq$20 %) of missing data, and therefore it might be considered beneficial for these databases to include the missing data. The subgroup analyses that applied the minimum number of reports criterion to the whole drug–event combination rather than within each stratum showed large increases in sensitivity but with some overall loss of precision for all variables in most databases. This approach is interesting and certainly the prospect of such improved sensitivity is attractive; however, the loss of precision is likely to lead to a higher absolute number of drug–event combinations being highlighted, which may be

unmanageable from a resource perspective in many organisations.

The practical implementation of subgroup analyses into routine first-pass signal detection may require some consideration to ensure that the absolute volume of work does not become unmanageable. Whilst stratified analyses result in a single combined disproportionality measure, subgroup analyses potentially will result in many different disproportionality measures from each subgroup. The implementation of subgroup analyses should therefore be such that a drug–event combination is highlighted for review without requiring the reviewer to evaluate each subgroup in detail as implemented in the vigiRank signal detection algorithm used at the Uppsala Monitoring Centre [12].

Although a range of databases of different sizes and characteristics were included in this study, the smallest still has 0.5 million reports. It is possible that databases with substantially fewer reports or products may produce different results to those observed in this study.

This study measured signal detection performance using overall sensitivity and precision. A further consideration with any signal detection system is the volume of drug–event combinations highlighted for review as most organisations will have a fixed resource to review these. Whilst sensitivity and precision can give some idea of likely relative volumes returned, i.e. a high sensitivity and low precision will produce high volumes of drug–event combinations, this study did not investigate how absolute volumes of drug–event combinations highlighted might vary across different approaches.

It was also beyond the scope of this study to evaluate whether different approaches using stratified, subgroup or crude analyses would capture the same ADRs and at a similar time.

The lack of a generally acceptable gold standard for determining ADRs is an issue for this type of study. In this study, SDRs were classified as true positives if they corresponded to ADRs labelled in the product information with those not labelled classified as false positives. This approach has been used in other studies [13, 17, 19]. This strategy will overestimate the number of false positives as some of these will turn out to be true and the reporting rate of known ADRs is likely to be different from those yet to be discovered. This should not undermine the comparison between the different approaches, however, since these are all compared using the same reference standard, but it is possible that the results may differ if the reference set was based on emerging safety signals rather than established ones as suggested by Norén et al. [27]. This study follows on from a previous PROTECT study comparing disproportionality methods [19] and uses the same methodology and reference standard. As part of this study a sensitivity analysis was conducted that restricted the reference set to ADRs identified in the post-authorisation phase and found that this did not change the findings substantively. This finding provides some reassurance in the context of the current study that the results would also be similar if based on ADRs identified post-authorisation.

A further limitation is that the study was conducted using a sample of products rather than all products in the database. This was done for practicality reasons and every effort was made to ensure that the study products represented a range of therapeutic areas and patient populations. This study did not consider the following additional areas of interest that may influence the choice of signal detection strategy: time to highlight an SDR, absolute volumes of SDRs, or any approach combining crude, stratified and/or subgroup analyses.

5 Conclusions

Subgroup analyses tended to perform better than stratified analyses in terms of sensitivity and precision in all spontaneous databases, although this was most evident in the two largest databases. Additionally, stratified analyses were not found to increase either sensitivity or precision beyond that associated with analytical artefacts of the stratified analysis and are unlikely therefore to provide added value. Subgroup analyses were shown to be beneficial in two large international databases with over 2 million reports with increases in both sensitivity and precision observed, although the observed increase in precision may also be an artefact of the subgrouping process rather than a true effect. Smaller datasets may need to consider a likely trade-off between increased precision with some loss of sensitivity if subgroup analysis was to replace a crude analysis. Covariates that showed the most promising results included age and region/country of origin, but it is likely that the choice of covariates for subgroup analyses will vary between different databases.

Acknowledgments The views expressed in this paper are those of the authors only and not of their respective organisations and do not reflect the official policy or position of the Innovative Medicines Initiative Joint Undertaking (IMI JU).

Compliance with Ethical Standards

Funding The research leading to these results was conducted as part of PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium; www.imi-protect.eu), which is a public–private partnership coordinated by the European Medicines Agency (EMA).
The PROTECT project has received support from the Innovative Medicines Initiative Joint Undertaking (IMI JU; www.imi.europa.eu) under Grant Agreement No 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007-2013) and companies of the

European Federation of Pharmaceutical Industries and Associations (EFPIA) in-kind contribution.

Conflict of interest Naashika Quarcoo and Jeffery Painter are employees of and hold shares in GlaxoSmithKline. Ramin Arani and Antoni Wisniewski are employees of and hold shares in AstraZeneca. Products from these companies were among those used to test the methodologies in this research. Suzie Seabroke, Gianmario Candore, Kristina Juhlin, Naashika Quarcoo, Antoni Wisniewski, Ramin Arani, Jeffery Painter, Philip Tregunno, Niklas Norén and Jim Slattery have no financial interest in any commercial signal detection software.

References

1. Rawlins MD. Spontaneous reporting of adverse drug reactions. II: uses. Br J Clin Pharmacol. 1988;26:7–11.
2. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6):483–6.
3. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315–21.
4. Van Puijenbroek EP, Diemont WL, Van Grootheest K. Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. Drug Saf. 2003;26(5):293–301.
5. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf. 2002;25(6):381–92.
6. Hauben M, Horn S, Reich L. Potential use of data-mining algorithms for the detection of 'surprise' adverse drug reactions. Drug Saf. 2007;30(2):143–55.
7. Hopstadius J, Norén GN, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. Drug Saf. 2008;31(11):1035–48.
8. Gogolak VV. The effect of backgrounds in safety analysis: the impact of comparison cases on what you see. Pharmacoepidemiol Drug Saf. 2003;12(3):249–52.
9. Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf. 2005;28(11):981–1007.
10. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf. 2003;12:559–74.
11. Hopstadius J, Norén GN. Robust discovery of local patterns: subsets and stratification in adverse drug reaction surveillance. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. New York: ACM; 2012. p. 265–73 (Abstract).
12. Caster O, Juhlin K, Watson S, Norén GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank. Drug Saf. 2014;37(8):617–28.
13. Van Holle L, Bauchau V. Optimization of a quantitative signal detection algorithm for spontaneous reports of adverse events post immunization. Pharmacoepidemiol Drug Saf. 2013;22(5):477–87.
14. Woo EJ, Ball R, Burwen DR, et al. Effects of stratification on data mining in the US Vaccine Adverse Event Reporting System (VAERS). Drug Saf. 2008;31(8):667–74.
15. Almenoff JS, LaCroix KK, Yuen NA, Fram D, DuMouchel W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf. 2006;29(10):875–87.
16. Zeinoun Z, Seifert H, Verstraeten T. Quantitative signal detection for vaccines: effects of stratification, background and masking on GlaxoSmithKline's spontaneous reports database. Hum Vaccin. 2009;5(9):599–607.
17. Grundmark B, Holmberg L, Garmo H, Zethelius B. Reducing the noise in signal detection of adverse drug reactions by standardizing the background: a pilot study on analyses of proportional reporting ratios-by-therapeutic area. Eur J Clin Pharmacol. 2014;70(5):627–35.
18. de Bie S, Verhamme KM, Straus SM, Stricker BH, Sturkenboom MC. Vaccine-based subgroup analysis in VigiBase: effect on sensitivity in paediatric signal detection. Drug Saf. 2012;35(4):335–46.
19. Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S, et al. Comparison of statistical signal detection methods within and across databases. Drug Saf. 2015;38(6):577–87.
20. van Puijenbroek EP, Bate A, Leufkens HGM, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11:3–10.
21. 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.
22. DuMouchel W. Bayesian data mining in large frequency tables with an application to the FDA spontaneous reporting system. Am Stat. 1999;53(3):177–90.
23. Robins J, Greenland S, Breslow N. A general estimator for the variance of the Mantel–Haenszel odds ratio. Am J Epidemiol. 1986;124:719–23.
24. Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in EudraVigilance post-authorisation data: a retrospective evaluation of the potential for earlier signalling. Drug Saf. 2010;33(6):475–87.
25. Bate A, Edwards IR, Lindquist M, Orre R. The authors' reply [letter]. Drug Saf. 2003;26(5):364–6.
26. Hopstadius J, Norén GN, Bate A, Edwards IR. Stratification for spontaneous report databases. Drug Saf. 2008;31(12):1145–7.
27. Norén GN, Caster O, Juhlin K, Lindquist M. Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance. Drug Saf. 2014;37:655–9.