ORIGINAL RESEARCH ARTICLE

Drug Safety 2008; 31 (11): 1035-1048
0114-5916/08/0011-1035/$48.00/0

Impact of Stratification on Adverse Drug Reaction Surveillance

Johan Hopstadius,[1] *G. Niklas Norén*,[1,2] *Andrew Bate*[1,3] and *I. Ralph Edwards*[1]

1 Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden
2 Department of Mathematics, Stockholm University, Stockholm, Sweden
3 School of Information Systems, Computing and Mathematics, Brunel University, London, UK

Abstract

Background and objectives: Automated screening for excessive adverse drug reaction (ADR) reporting rates has proven useful as a tool to direct clinical review in large-scale drug safety signal detection. Some measures of disproportionality can be adjusted to eliminate any undue influence on the ADR reporting rate of covariates, such as patient age or country of origin, by using a weighted average of stratum-specific measures of disproportionality. Arguments have been made in favour of routine adjustment for a set of common potential confounders using stratification. The aim of this paper is to investigate the impact of using adjusted observed-to-expected ratios, as implemented for the Empirical Bayes Geometric Mean (EBGM) and the information component (IC) measures of disproportionality, for first-pass analysis of the WHO database.

Methods: A simulation study was carried out to investigate the impact of simultaneous adjustment for several potential confounders based on stratification. Comparison between crude and adjusted observed-to-expected ratios were made based on random allocation of reports to a set of strata with a realistic distribution of stratum sizes. In a separate study, differences between the crude IC value and IC values adjusted for (combinations of) patient sex, age group, reporting quarter and country of origin, with respect to their concordance with a literature comparison were analysed. Comparison was made to the impact on signal detection performance of a triage criterion requiring reports from at least two countries before a drug-ADR pair was highlighted for clinical review.

Results: The simulation study demonstrated a clear tendency of the adjusted observed-to-expected ratio to spurious (and considerable) underestimation relative to the crude one, in the presence of any very small strata in a stratified database. With carefully implemented stratification that did not yield any very small strata, this tendency could be avoided. Routine adjustment for potential confounders improved signal detection performance relative to the literature comparison, but the magnitude of the improvement was modest. The improvement from the triage criterion was more considerable.

Discussion and conclusions: Our results indicate that first-pass screening based on observed-to-expected ratios adjusted with stratification may lead to missed signals in ADR surveillance, unless very small strata are avoided. In addition, the improvement in signal detection performance due to routine adjustment for a set

of common confounders appears to be smaller than previously assumed. Other approaches to improving signal detection performance such as the development of refined triage criteria may be more promising areas for future research.

Background

The analysis of individual case safety reports (ICSRs) related to suspected adverse drug reaction (ADR) incidents in clinical practice remains the most important method to discover unexpected adverse effects from drug substances that are already on the market.[1] ICSR submission is also sometimes referred to as spontaneous reporting. The Uppsala Monitoring Centre (UMC) maintains and analyses the world's largest database of international ADR reports on behalf of the WHO Programme for International Drug Monitoring. The main use of the WHO database is to raise hypotheses about suspected ADRs, also referred to as signals. The WHO definition of a signal is:

"Reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously. Usually more than a single report is required to generate a signal, depending upon the seriousness of the event and the quality of the information."[2]

While careful clinical review of ICSRs remains a critical and irreplaceable step of the signal detection process, knowledge discovery methods are becoming increasingly important to filter the massive amounts of data for interesting associations and for pattern recognition to identify unexpected characteristics of groups of ICSRs.[3,4] A knowledge discovery framework for hypothesis generation in ADR surveillance has been in routine use as part of the UMC signal detection process since 1998[5] and has helped highlight several important drug safety problems later published in the medical literature.[6,7] Central to the initial phase of UMC's knowledge discovery framework is the information component (IC) measure of disproportionality.[3,8-10] The IC is based on an observed-to-expected ratio that contrasts the relative reporting rate of an ADR given a particular drug to the overall relative reporting rate of the ADR in the database.[3] It is a so-called shrinkage measure that tends to the baseline value of 0, until a large enough number of reports on a particular drug-ADR pair have accumulated. This decreases the risk of highlighting chance disproportionality. The formulae for computing the IC are complicated, but well approximated by equation 1:[11]

$$IC = \log_2 \frac{O_{xy} + 1/2}{E_{xy} + 1/2} \quad \text{(Eq. 1)}$$

where O_{xy} is the observed number of reports on the ADR and the drug together and E_{xy} is the expected number of such reports conditional on the overall relative reporting rate of the ADR in the database and the total number of reports on the drug in the database. Accurate formulae for 95% credibility intervals for both the crude and the adjusted IC are available.[9] The lower limit of the 95% credibility interval is referred to as IC_{025}, and is the standard measure used to screen the WHO database for excessive ADR relative reporting rates. The Empirical Bayes Geometric Mean (EBGM) is an alternative shrinkage measure based on the same observed-to-expected ratio as the IC and with similar properties.[12,13]

Disproportionality analysis is only the first step in the UMC knowledge discovery process. Triage (prioritization) algorithms have been designed to focus attention on the drug-ADR pairs for which follow-up is most urgent and to highlight drug-ADR pairs also in the absence of disproportional reporting rates.[14] All highlighted drug-ADR pairs are reviewed by subject matter experts to identify issues worthy of general communication.

An important obstacle in the analysis of any observational dataset is the possible presence of measured or unmeasured confounders – covariates that distort the quantitative relationships under study. Downward confounders are covariates that

mask true intrinsic associations. An example in the WHO database is the negative crude IC value for the association between tamoxifen and impotence.[15] Tamoxifen is primarily used in women and impotence primarily affects men. The expected number of reports on tamoxifen and impotence, taking this into account, is considerably less than the crude expected number based on the number of reports on tamoxifen and the number of reports on impotence in the database. This leads to the lower crude than adjusted IC value. In contrast, upward confounders are covariates that result in false apparent associations between events of interest. An example in the WHO database is the positive crude IC value for the association between *Haemophilus influenzae* type B vaccine and fever convulsions.[15] This is due to the lower rate of fever convulsions in the database as a whole than in the young children to whom the vaccine is primarily given, which leads to an underestimated expected number of reports.

Adjustment of the IC for potential confounders is possible by stratifying the database into subgroups based on each report's value for the suspected confounder, and replacing the overall expected count in the equation for computing the IC value (equation 1) by a sum over stratum-specific expected counts.[9,12] This may eliminate the unwanted impact of the confounder adjusted for. However, when there is considerable variation between strata, no overall measure is appropriate. Confounding can only be evaluated in the absence of effect modification.

The first-pass screening for disproportional reporting rates in the WHO database has traditionally been based on crude IC values.[3,8-10] The advantages of this are transparency and computational efficiency. As a complement, stratum-specific variation is also considered.[10] Arguments have been made in favour of routine adjustment based on stratification.[12] However, although a range of comparative studies to investigate the performance of different data-mining algorithms for ICSRs have been carried out,[16-19] few have explicitly considered the impact of routine adjustment for potential confounders based on stratification. One exception is the study by Almenoff et al.,[20] who found that the number of associations highlighted based on the EBGM disproportionality measure decreased with stratification. However, this study did not investigate to what extent stratification selectively eliminated the false positives or also led to missed true signals. Indeed, in an earlier study of the WHO database based on a fixed threshold at $IC_{025}>0$, simultaneous adjustment for patient age, patient sex, reporting quarter and country of origin increased specificity at the expense of decreased sensitivity relative to a literature reference.[15]

Aim

The aim of this paper is to determine to what extent routine adjustment for a set of potential confounders based on stratification improves first-pass screening of the WHO database, and to investigate whether stratification that is too fine may have a negative impact on the adjusted observed-to-expected ratio.

Methods

Data Source

The studies presented in this paper are based on data from the WHO database as of 30 September 2006. At that point in time, the database contained more than 3.7 million reports of suspected ADRs – the first of which dated back to 1967. All in all, over 700 000 unique drug-ADR pairs had been reported together at least once in the database. More than 80 countries participated in the programme and the current inflow of new reports was in the order of 200 000 per year.

Covariates

There is a wide range of potential confounders in ADR surveillance, which cannot all be accounted for routinely in first-pass disproportionality analysis. A feasible approach is to routinely adjust for a more restricted selection of covariates expected to be the most important confounders in general. In observed-to-expected ratio-based disproportionality analysis with the EBGM, patient age, patient sex

and time of reporting are routinely adjusted for.[13,14] For international ADR surveillance, it has been argued that country of origin is also an important potential confounder.[21] As a consequence, we limit the following studies to these four covariates. Although concomitant medication is likely to be one of the most important types of confounders in ADR surveillance, we do not consider it further in the context of this paper. It has earlier been demonstrated that routine simultaneous adjustment for all possible combinations of concomitant medication using stratification leads quickly to a large number of very small strata and to severe overstratification where the adjusted observed-to-expected ratio tends to 1, even for the most excessive crude relative reporting rates.[15] As a consequence, routine adjustment for co-medication would require a different approach, such as shrinkage regression.[22]

A Simulation Study of the Impact of Overstratification

The adjusted observed-to-expected ratio discussed in the Background section is similar to the Mantel-Haenszel odds ratio in that both can be seen as weighted averages of stratum-specific measures of association. One advantage of the Mantel-Haenszel odds ratio[23] is that it is also accurate for sparse data.[24] However, it is unclear whether the same property holds for the adjusted observed-to-expected ratio. We wanted to investigate the impact of adjusting the observed-to-expected ratio for several potential confounders at once using stratification.

For this purpose, we partitioned the WHO database based on a cross-classification of reports into subcategories based on patient sex, patient age group, country of origin and time of reporting. We used the standard age groups already available in the WHO database (age unspecified; 0–1 month; 2 months–4 years; 5–11 years; 12–16 years; 17–69 years; and 70+ years), we divided reporting time into quarterly (3-month) intervals, in correspondence with the intervals in which the WHO database is routinely screened, and we used separate strata for each country of origin. This produced a stratification where the 3 704 938 case reports in the WHO database as of 30 September 2006 were partitioned into 34 756 different strata of varying sizes. To eliminate any true confounding due to the four covariates by which data were originally stratified, all reports were randomly re-allocated across strata while keeping the resulting numbers of reports per stratum fixed. Crude and adjusted observed-to-expected ratios were calculated based on this simulated stratification. As the allocation of reports to strata was random, the simulated stratification should have no true impact on the observed-to-expected ratio, and any systematic differences between the crude and adjusted observed-to-expected ratios would indicate a potential problem with the adjusted observed-to-expected ratio. To investigate the impact of shrinkage on the potential vulnerability of the observed-to-expected ratio to overstratification, we also compared crude and adjusted IC values based on the simulated stratification. In order to show whether any sensitivity to overstratification is a property of the observed-to-expected ratio on which the IC and the EBMG are based, and not an artefact of the specific shrinkage used, the simulation study considered both shrunk and unshrunk measures. The mathematical details that motivate why vulnerability for overstratification is a property of the observed-to-expected ratio are given in the Appendix.

A Database Study of the Impact of Different Potential Confounders

Under the assumption that overstratification may be a problem for measures of disproportionality based on the observed-to-expected ratio, it is important to find ways to reduce the stratification granularity. One approach is to decrease the number of potential confounders adjusted for. The main question then is which of our four potential confounders have the most considerable overall impact on drug-ADR disproportionality in the WHO database. To be able to determine this, we investigated to what extent adjustment by different sets of covariates led to differences between the crude and adjusted IC values. We also studied to what extent adjustment for different sets of confounders improved signal detection performance in ADR surveillance. The

lack of a general 'gold standard' for evaluating signal detection performance has been discussed elsewhere.[25] We used as a reference here the overview of first published case reports of suspected ADRs in the international medical literature provided in the Wolters Kluwer Health | Adis publication *Reactions Weekly* (http://www.reactions.adisonline.com). This literature reference has also been used in an earlier evaluation of signal detection performance on the WHO database.[5] The suspected ADR incidents referred to in *Reactions Weekly* correspond to first reports of suspected ADRs in the peer reviewed international literature and therefore have been suggestive enough to be considered worthy of follow-up. They should represent a fair overview of emerging knowledge related to suspected ADRs on an international level and, as such, constitute a reasonable reference for evaluating signal detection performance in the WHO database. A practical benefit of the *Reactions Weekly* information is that it is routinely encoded in the WHO database and is readily available from within the system. As the IC analysis methodology has been designed not to highlight drug-ADR pairs for clinical review based on less than three reports, the scope of the study was limited to drug-ADR pairs with three or more reports in the WHO database.

We limited our investigation to the four covariates described in the Covariates section: patient age; patient sex; time of reporting; and country of origin. In order to generally reduce the risk of overstratification, we used more coarse categories than those considered in the Covariates section. Patient age was categorized according to World Organisation of National Colleges Acadamies and Academic Associations of General Practitioners/Family Physicians (WONCA) standard age groups:[26] unspecified; <1 year, 1–4 years; 5–14 years; 15–24 years; 25–44 years; 45–64 years; 65–74 years; and 75+ years. The advantage of this coarser categorization is mainly that some very small strata for the youngest patients in the standard age groups of the WHO database were avoided. Time of reporting was divided into intervals consisting of several years: 1967–75; 1976–80; 1981–5; 1986–90; 1991–3; 1994–6; 1997–9; 2000–2; and 2003–6 (the interval length diminishing with calendar time as the number of reports per year increases). In addition, countries with <100 reports in total were grouped in a single small country stratum to avoid overstratification by country of origin. To avoid small strata when combining time of reporting and country of origin, time periods were automatically merged for countries with little or irregular reporting until each country-time stratum contained at least 100 reports.

To study the signal detection performance of the different IC_{025} values independently of the number of highlighted drug-ADR pairs, we used precision-recall graphs. In this context, precision is defined as the proportion of highlighted drug-ADR pairs that were included in the literature reference and recall is defined as the proportion of drug-ADR pairs included in the literature reference (with at least three reports in the WHO database) that were actually highlighted. A precision-recall graph plots precision versus recall at varying thresholds, and in that it is similar to a receiver-operating characteristic (ROC) curve.[27] However, ROC curves are not suitable for applications with a massive number of true negatives (such as ADR surveillance where most reported drug-ADR pairs do not correspond to new drug safety problems) for which the proportion of true negatives among non-highlighted cases (the specificity) will be near 1 for all relevant thresholds.

Results

Simulation Study

Figure 1 plots adjusted versus crude log observed-to-expected ratios for all drug-ADR pairs that co-occur on at least three reports in the WHO database, based on the simulated stratification described in the methods section. Under the random re-allocation of reports across strata, the expected value of the adjusted observed-to-expected ratio should be the same as that of the crude observed-to-expected ratio, and the scatter plot should have been roughly diagonal, with some random variation around. In contrast, figure 1 displays a clear tendency of the adjusted

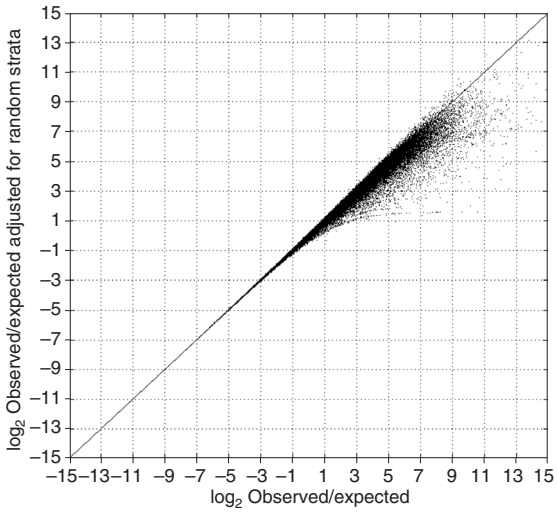

Fig. 1. Adjusted vs crude observed-to-expected ratios for a random allocation of reports to strata with the same size distribution as the WHO database when simultaneously stratified by patient sex, patient age (seven groups), country of origin and time of reporting (in quarterly intervals).

observed-to-expected ratio to spuriously fall below that of the crude (in particular for large crude observed-to-expected ratios). For most combinations, the difference between the crude and the adjusted observed-to-expected ratios is only minor as seen in figure 2, which is a histogram of the difference between the crude and the adjusted observed-to-expected ratios. Altogether, there are 3044 drug-ADR pairs for which the logarithm of the adjusted observed-to-expected ratio falls more than 1 unit below that of the crude, and 5 drug-ADR pairs for which the logarithm of the adjusted observed-to-expected ratio exceeds that of the crude by more than 1 unit. Figures 3 and 4 display the same graphs based on IC values. Shrinkage reduces but does not eliminate the problem with spurious underestimation. There are 565 drug-ADR pairs for which the adjusted IC falls more than 1 unit below that of the crude. These are considerable differences in that they correspond to an overestimation of >100% or an underestimation of >50% of the underlying observed-to-expected ratio. For 2127 drug-ADR pairs, the crude but not the adjusted IC_{025} exceeds 0, and for 1039 drug-ADR pairs, the adjusted but not the crude IC_{025} exceeds 0. These represent discrepancies in the signals of disproportional reporting highlighted with the two measures and therefore emerging drug safety issues that could be missed. At least for some drug-ADR pairs, the spurious underestimation is considerable and likely to outweigh any beneficial impact of actual confounder elimination in a real-world analysis.

The numbers quoted in the previous paragraph are from a single realization of the simulation experiment. To confirm that variability in the simulation results is not excessive, we repeated the experiment five times. All quoted results were stable over these repeated simulations. For example, the number of drug-ADR pairs for which the absolute value of the difference between the crude and adjusted log observed-to-expected ratio exceeds 1 varied between 3047 and 3177. For the IC, the corresponding range was 505–614.

Database Study

Figure 5 illustrates the general impact of the four different covariates on drug-ADR disproportionality in the WHO database. The scatter plots compare crude and adjusted IC values for all drug-ADR associations with at least three reports in the WHO database. The more the point clouds deviate from a narrow diagonal line, the greater are the discrepancies between crude and adjusted IC values. Clearly,

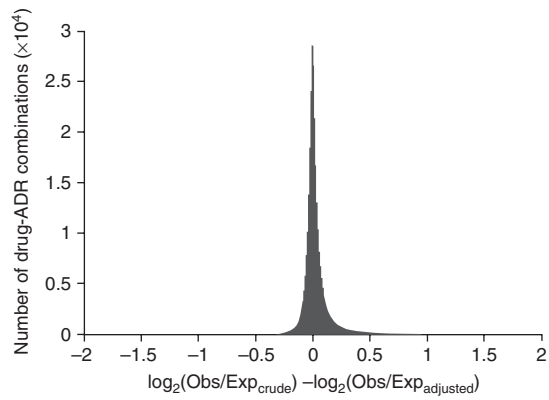

Fig. 2. Histogram displaying the difference between the crude and the adjusted log observed-to-expected ratios. This corresponds to the spread perpendicular to the diagonal of figure 1. **ADR** = adverse drug reaction.

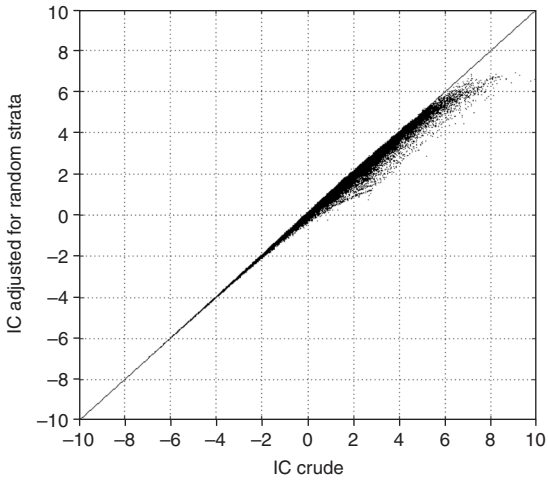

Fig. 3. Adjusted vs crude information component (IC) for a random allocation of reports to strata with the same size distribution as the WHO database when simultaneously stratified by patient sex, patient age (seven groups), country of origin and time of reporting (in quarterly intervals).

the impact of patient sex is considerably less than that of the other three covariates.

Figure 6 displays precision-recall graphs based on the crude IC_{025} value and based on IC_{025} values adjusted for each of the four covariates. Figure 7 is an enlarged version of figure 6, zoomed in around the area of the precision recall-graph corresponding to the current threshold (marked with an x) which is used routinely in the WHO database to highlight drug-ADR pairs for clinical review. Clearly, the adjustment by country of origin and year of reporting, respectively, led to the most considerable improvements in performance relative to the literature reference. Based on the major performance improvement of separate adjustment for country of origin and time of reporting, we wanted to investigate their combined effect. Figure 8 displays the precision-recall graph for simultaneous adjustment by time of reporting and country of origin. Precision-recall graphs based on crude IC_{025} values and IC_{025} values adjusted for country of origin or time of reporting are included for comparison. Figure 9 is an enlarged version of figure 8, zoomed in around the area of the precision recall-graph corresponding to the current threshold (marked with an x). Clearly, simultaneous adjustment for the two covariates further improves performance relative to separate adjustment for either covariate. To demonstrate that the IC simultaneously adjusted for country of origin and time of reporting does not suffer from overstratification, figure 10 displays a comparison of adjusted and crude observed-to-expected ratios based on randomization of reports across strata as in the simulation described in the methods section. Figure 11 demonstrates that the spurious differences between adjusted and crude observed-to-expected ratios are minimal with this stratification. We expect this is due to the avoidance of any strata with fewer than 100 reports in this stratification approach.

To put into perspective the magnitude of the performance improvement as a result of simultaneous adjustment of the IC_{025} for time of reporting and country of origin, figure 12 displays precision-recall graphs for the crude and adjusted IC_{025} values together with precision-recall graphs for the crude and adjusted IC_{025} values when combined with a triage criterion requiring reports from more than one country to highlight a drug-ADR pair for clinical review (as originally proposed by Ståhl et al.[14]). While the adjustment for time of reporting and country of origin does improve performance both with and without the triage criterion, the magnitude of the improvement is rather small relative to that due to the triage criterion. This supports the comment by

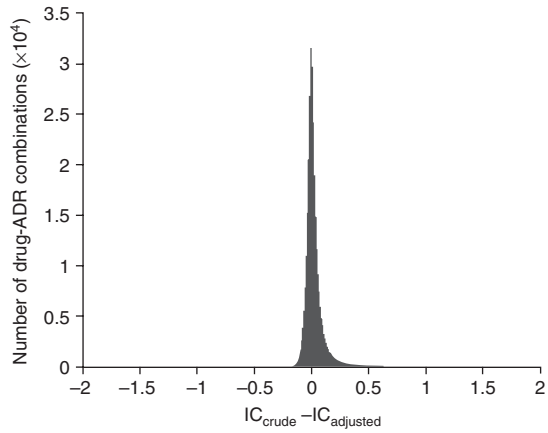

Fig. 4. Histogram displaying the difference between the crude and the adjusted information component (IC). This corresponds to the spread perpendicular to the diagonal of figure 3. **ADR** = adverse drug reaction.

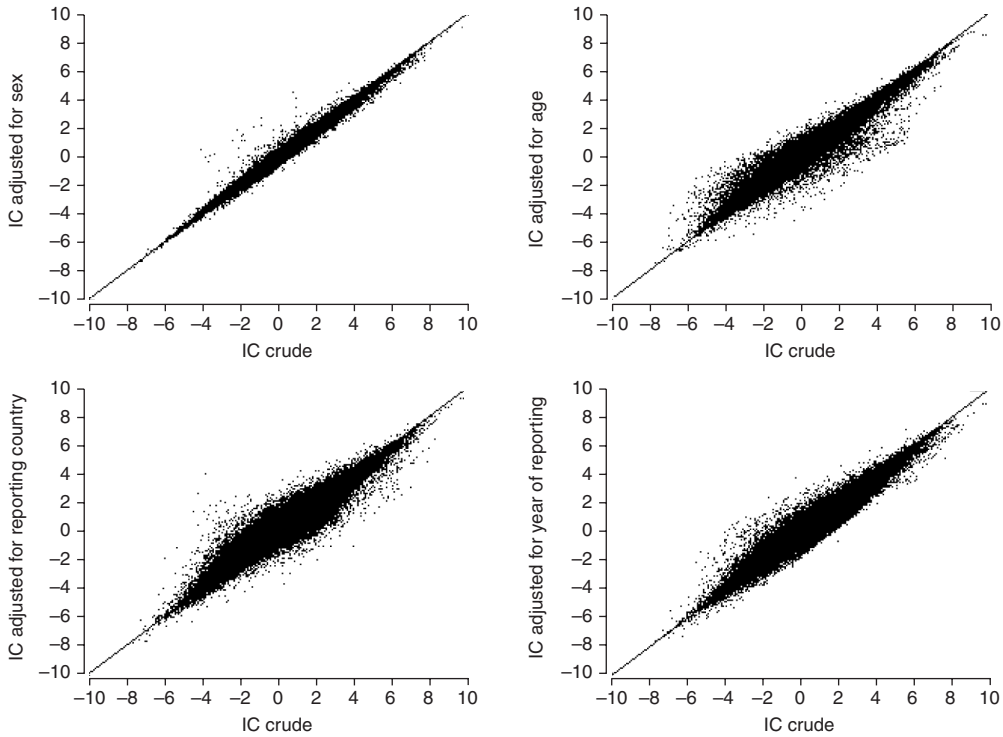

Fig. 5. Scatter plots of adjusted vs crude information component (IC) values for each of the four covariates considered in this paper, based on all drug-adverse drug reaction combinations with at least three reports in the WHO database.

Bate et al.[28] that routine adjustment for a group of common confounders is perhaps not among the most important possible improvements to first-pass exploratory analysis of ICSRs.

Summary

Observed-to-expected ratios adjusted based on stratification must be carefully monitored for over-stratification: the presence of any small strata may lead to spurious underestimation. Careful adjustment based on stratification by time interval of reporting and country of origin did improve signal detection performance in the WHO database relative to a literature reference. The general impact of adjusting based on stratification by patient age and sex was less. In contrast, the performance improvement due to a triage criterion requiring reports from more than one country to highlight a drug-ADR pair for clinical review was considerably greater than that from routine adjustment for any potential confounder(s) considered in this study.

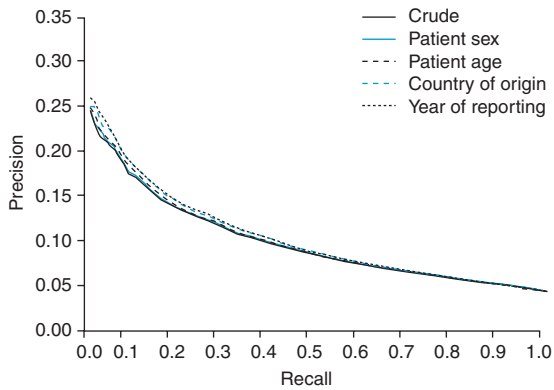

Fig. 6. Graphs indicating the precision and recall relative to the literature reference at varying thresholds, for four different IC_{025} values: crude; adjusted for patient sex; patient age; time of reporting; and country of origin.

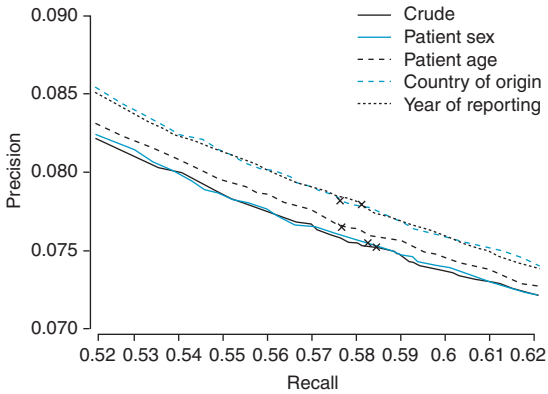

Fig. 7. Enlarged version of figure 6 around the area corresponding to a threshold at IC_{025} = 0. The exact points on the graphs corresponding to the threshold are marked with crosses.

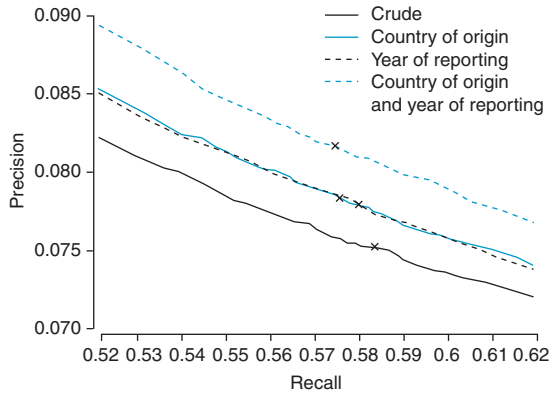

Fig. 9. Enlarged version of figure 8 around the area corresponding to a threshold at IC_{025} = 0. The exact points on the graphs corresponding to the threshold are marked with crosses.

Discussion

Routine adjustment for a standard set of potential confounders based on stratification has been assumed to be a useful approach in first-pass screening of spontaneous reports for disproportional reporting rates. The higher specificity of the adjusted EBGM compared with the unadjusted EBGM observed in some empirical studies has been interpreted as evidence of successful confounder elimination;[20,29] however, our results indicate that spurious underestimation is an alternative explanation. In this paper, we have demonstrated that observed-to-expected ratios are vulnerable to overstratification. Even stratifying by a moderate number of covariates may lead to unreliable adjusted observed-to-expected ratios. In the WHO database, stratification by a dummy covariate with the size and number of strata as based on simultaneous adjustment by patient age, sex, reporting quarter and country of origin, led to considerable, spurious underestimation of both the adjusted observed-to-expected ratio and the adjusted IC. Because of this sensitivity of the adjusted observed-to-expected ratio to overstratification, strategies for routine adjustment of the IC or the EBGM

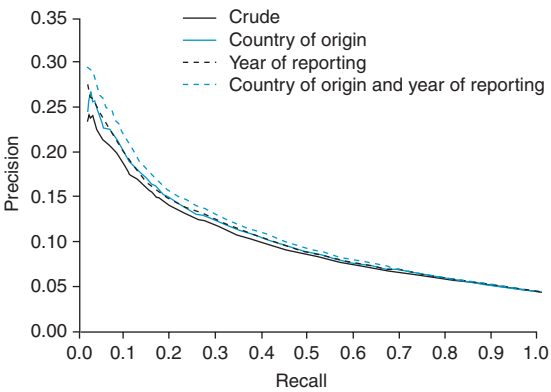

Fig. 8. Graphs indicating the precision and recall relative to the literature reference at varying thresholds, for four different IC_{025} values: crude; adjusted for reporting time interval; adjusted for country of origin; and simultaneously adjusted for reporting time interval and country of origin.

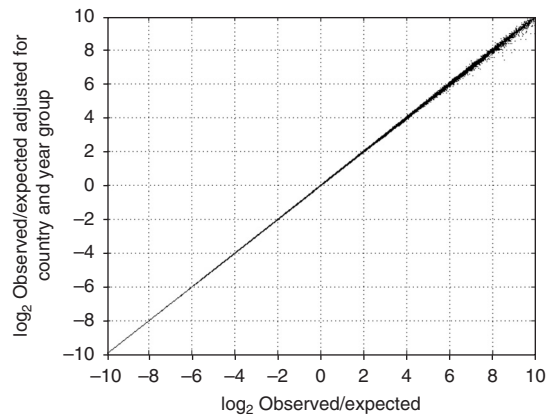

Fig. 10. Adjusted vs crude $\log_2$ observed-to-expected ratios, for a random allocation of case reports to strata with the same size distribution as the WHO database when simultaneously stratified by reporting time interval and country of origin.

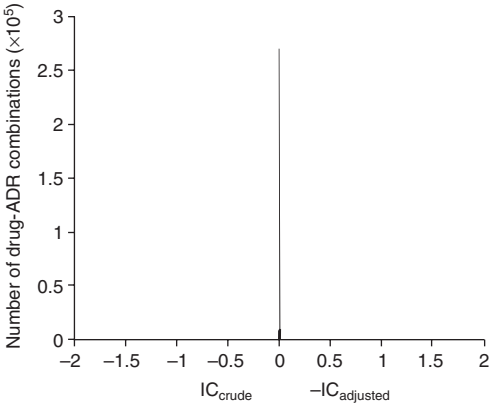

Fig. 11. Histogram displaying the difference between the crude and the adjusted log observed-to-expected ratios. This corresponds to the spread perpendicular to the diagonal of figure 10. **ADR** = adverse drug reaction.

based on stratification must be carefully designed so that they do not create any very small strata.

In our simulation study, stratifying by the dummy covariate did not confound any drug-ADR pairs. Thus, the adjusted observed-to-expected ratio could at best perform as well as the crude ratio. In real-world data analysis, the adjusted observed-to-expected ratio may eliminate the impact of some confounders. However, the spurious underestimation in our simulation was of such magnitude that it cannot be assumed ignorable relative to any beneficial impact of successful elimination of confounding.

While the unique international coverage of the WHO database makes it unusually heterogeneous, overstratification is an important concern also for more homogeneous datasets. The observed-to-expected ratio is sensitive to the presence of any small strata, and given the more limited scope of company ADR surveillance systems, even a partitioning of reports into the approximately 900 strata used in routine adjustment of the EBGM[29,30] will produce some very small strata. In addition, continuous ADR surveillance may lead to a situation where strata related to the most recent time period contain very few reports, regardless of overall dataset size. Clearly, careful adjustment of the observed-to-expected ratio is important in any analysis of ICSRs.

We have chosen to use the first reports information in *Reactions Weekly* as reference for evaluating signal detection performance. Other choices are certainly possible, and may lead to slightly different results. However, we do not expect there to be systematic variations to the extent that our conclusions, with respect to the relative importance of routine adjustment for potential confounders, would be invalidated by a different choice of reference. The sensitivity of the observed-to-expected ratio to overstratification was demonstrated independently of the selected gold standard.

An important advantage of adjustment based on stratification by country of origin and time of reporting is that it usually yields transparent adjusted measures that correspond closely to stratum-specific measures based on those time periods and countries in which both the drug and the ADR have been reported. In addition, there are no missing data for these covariates on any reports in the WHO database, and confounding by country of origin and time of reporting usually corresponds to aspects of data collection that may be more difficult to identify in clinical review than confounding by patient characteristics. At the same time, the lack of overall improvement from adjusting for patient sex and age does not imply that these covariates are not sometimes important confounders. Some examples of drug-ADR pairs with great discrepancies between crude IC values and IC values adjusted for patient

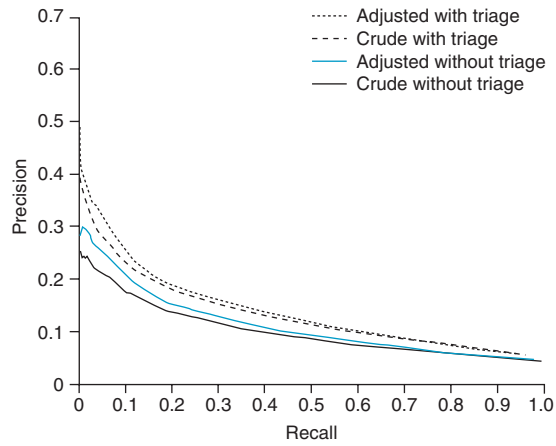

Fig. 12. Precision-recall graphs for the crude IC_{025} and the IC_{025} simultaneously adjusted for reporting time interval and country of origin, with and without a triage criterion requiring reports from more than one country.

sex or age have previously been presented.[15,20] However, the majority of these examples correspond to simple stratum-specific effects, such as ADRs that by definition only ever affect male patients, and they are usually easy to identify in the clinical review.

Routine strategies to screen for stratum-specific disproportionality are clearly an important complement to screening for overall disproportionality. Moreover, if first-pass screening is based on unstratified disproportionality analysis, methods for highlighting suspected confounding are also important. A standard approach to data-driven confounder identification is to look for changes in a measure due to adjustment by different covariates.[15] A more open-ended approach may be to, as a first step, use disproportionality analysis to identify covariates that are disproportionally reported with both the drug and the ADR. This is an important area for future research.

This paper has focused on the observed-to-expected ratio, primarily because it is the basis for routine surveillance of the WHO database. While extreme overstratification would have a negative impact also on adjusted proportional reporting ratios (PRRs)[4] and Mantel-Haenszel adjusted reporting odds ratios (RORs),[31] they appear to be far less sensitive to the presence of small numbers of very small strata. Their main limitation in the context of ADR surveillance is that they are vulnerable to spurious associations based on very few reports and require additional filters to be useful in first-pass screening.[4] Regression-based adjustment for potential confounders is an alternative to the post-stratification approach discussed here. Its main advantages are that it allows for direct adjustment by numerical covariates and that several covariates can be adjusted for simultaneously without accounting for potential interaction between confounders. However, regression-based adjustment may not be considered transparent and can still be sensitive to overstratification.[32] Adjustment for suspected confounders based on propensity scores may be a useful approach to handle multiple confounders, since propensity scores project reports on to a small number of strata. However, propensity score-based adjustment is a form of automated confounder selection strategy that must be applied separately to each drug. Changes in a database over time could lead to changes in the propensity score projections that, in turn, induce unexpected changes in the adjusted measures. In the context of ADR surveillance, adjustment for suspected confounders based on propensity scores is therefore perhaps more appropriate for hypothesis refinement and strengthening.

Conclusions

It is clear that if a measure of disproportionality based on the observed-to-expected ratio is routinely adjusted for a set of potential confounders, the user must beware that the reduction in false positives may come at a cost of missed signals. Indeed, given the rather modest performance improvement from routine adjustment for potential confounders observed in this study, it remains an open question regarding how much effort should be spent on implementing and improving strategies for routine adjustment in ADR surveillance. Our results indicate that resources may be better spent on research to improve triage algorithms for signal selection and follow-up.

Acknowledgements

The authors are indebted to all the national centres that make up the WHO Programme for International Drug Monitoring and contribute case reports to the WHO database. However, the opinions and conclusions are not necessarily those of the various centres or of the WHO. The authors would also like to thank Professor Rolf Sundberg for helpful comments on an earlier draft of this article.

No sources of funding were used to assist in the preparation of this study. The authors have no conflicts of interest that are directly relevant to the content of this study.

Appendix

Further Mathematical Detail Relating to Effects of Overstratification

Our simulation study provides empirical evidence that the adjusted observed-to-expected ratio tends to spurious underestimation in the presence of

any very small strata. In this appendix, we provide some mathematical details for the interested reader who wants to understand the background to this behaviour. Consider the following 2 × 2 contingency table for the reporting of drug *x* and ADR *y* (equation 2):

	y	not *y*	
x	*a*	*b*	*a* + *b*
not *x*	*c*	*d*	
	a + *c*		*a* + *b* + *c* + *d*

(Eq. 2)

Conditional on the marginal number of reports on x (a + b), the marginal number of reports on y (a + c) and the total number of reports (a + b + c + d), the observed-to-expected ratio (OE) is (equation 3):

$$OE = \frac{a/(a+b)}{(a+c)/(a+b+c+d)}$$

(Eq. 3)

where the observed and the expected number of reports on *x* and *y* together can be expressed as (equations 4 and 5):

$$O_{xy} = a$$

(Eq. 4)

$$E_{xy} = \frac{(a+b)(a+c)}{a+b+c+d}$$

(Eq. 5)

A weighted average of the stratum-specific observed-to-expected ratios, with weights equal to the stratum-specific expected number of reports is (equation 6):

$$OE_{adj} = \frac{\Sigma_z \frac{O^z_{xy}}{E^z_{xy}} \cdot E^z_{xy}}{\Sigma_z E^z_{xy}} = \frac{O_{xy}}{\Sigma_z E^z_{xy}}$$

(Eq. 6)

where the expected number of reports in each stratum z is (equation 7):

$$E^z_{xy} = \frac{(a_z + b_z)(a_z + c_z)}{a_z + b_z + c_z + d_z}$$

(Eq. 7)

This adjusted observed-to-expected ratio was first proposed for the EBGM by DuMouchel,[12] and later adopted for the IC.[9]

In practice, O^z_{xy} and E^z_{xy} will be zero in any stratum z that contains no reports on *x* or no reports on *y*. Therefore, in some cases, the adjusted observed-to-expected ratio corresponds closely to the crude observed-to-expected ratio for a specific subset of the database. For example, adjustment by country of origin and time of reporting will restrict the basis for the expected number of reports on a drug-ADR pair to those countries and time periods in which both the drug and the ADR have been reported. The advantage of this is that it may eliminate any undue impact of geographical and temporal variations in reporting rates. The drawback is that for rare drugs and ADRs, the adjusted E_{xy} may be based on a very small subset of the database, and thereby be more sensitive to sampling variability than the unadjusted E_{xy}.

A closer investigation reveals that the presence of any very small strata has a negative impact on the adjusted observed-to-expected ratio. Consider the adjusted observed-to-expected ratio expressed in terms of the counts in the 2 × 2 contingency table (equation 8):

$$OE_{adj} = \frac{a}{\Sigma \frac{(a_z + b_z)(a_z + c_z)}{a_z + b_z + c_z + d_z}}$$

(Eq. 8)

The adjusted observed-to-expected ratio is strongly affected by single report strata in which the drug and the ADR co-occur (i.e. with $a_z = 1$ and $b_z = c_z = d_z = 0$). In such a stratum, both the observed and the expected counts are 1. Conditional on the stratum-specific marginals, there is no randomness left in a_z for such degenerated tables, and the observed number of reports will always equal the expected. Moreover, given that the stratum-specific observed-to-expected ratio is based on one single report, it is given an unreasonably large weight in the weighted average (equation 8). If the overall crude expected count is small (as is often the case in ADR surveillance), the dampening effect on the overall observed-to-expected ratio will be considerable.

Counter-intuitively, if placed in a separate stratum, an additional ICSR on a particular drug-ADR pair will moderate the original observed-to-expected ratio towards 1. This problem is not specific to single report strata. Consider a stratum with originally four reports such as $a_z = b_z = c_z = 0$ and $d_z = 4$. If a report on the drug-ADR pair of interest is placed in this stratum so that b_z and c_z remain 0, d_z remains 4, but a_z increases to 1, then E_{xy}^{z} increases from 0 to 0.2. If the original overall observed count was 3 and the overall adjusted expected count was 0.01, then the overall observed-to-expected ratio decreases from 3/0.01 = 300 to 4/0.21 ≈19, even though the observed count has increased by 1 and the expected count by only 0.2. Similar examples can be found where added reports on a drug-ADR pair reduce the observed-to-expected ratio, even if placed in a stratum where any of a_z, b_z, c_z and d_z are non-zero. Shrinkage will reduce but not eliminate this problem.

References

1. Rawlins MD. Spontaneous reporting of adverse drug reactions. II: uses. Br J Clin Pharmacol 1988; 26 (1): 7-11
2. Edwards IR, Biriell C. Harmonisation in pharmacovigilance. Drug Saf 1994; 10 (2): 93-102
3. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998; 54 (4): 315-21
4. Evans SJ, 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
5. Lindquist M, Edwards IR, Bate A. From association to alert: a revised approach to international signal analysis. Pharmacoepidemiol Drug Saf 1999; 8 (S1): S15-25
6. Sanz EJ, De-las-Cuevas C, Kiuru A, et al. Selective serotonin reuptake inhibitors in pregnant women and neonatal withdrawal syndrome: a database analysis. Lancet 2005; 365 (9458): 482-7
7. Coulter DM, Bate A, Meyboom RH, et al. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. BMJ 2001; 322 (7296): 1207-9
8. Bate A, Lindquist M, Edwards IR. A data mining approach for signal detection and analysis. Drug Saf 2002; 25 (6): 393-7
9. Norén GN, Bate A, Orre R, et al. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med 2006; 25 (21): 3740-57
10. Orre R, Lansner A, Bate A, et al. Bayesian neural networks with confidence estimations applied to data mining. Comput Stat Data Anal 2000; 34 (8): 473-93
11. Norén GN. Statistical methods for knowledge discovery in adverse drug reaction surveillance [doctoral thesis]. Stockholm: Faculty of Science, Department of Mathematics, Stockholm University, 2007
12. 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
13. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf 2003; 12 (7): 559-74
14. Ståhl M, Lindquist M, Edwards IR, et al. Introducing triage logic as a new strategy for the detection of signals in the WHO Drug Monitoring Database. Pharmacoepidemiol Drug Saf 2004; 13 (6): 355-63
15. Hopstadius J. Methods to control for confounding variables in screening for association in the WHO drug safety database [master's thesis]. Uppsala: Department of Mathematics, Uppsala University, 2006
16. Roux E, Thiessard F, Fourrier A, et al. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. IEEE Trans Inf Technol Biomed 2005; 9 (4): 518-27
17. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. Drug Saf 2004; 27 (10): 735-44
18. Lindquist M, Stahl M, Bate A, et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO International Database. Drug Saf 2000; 23 (6): 533-42
19. van Puijenbroek EP, Bate A, Leufkens HG, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002; 11 (1): 3-10
20. Almenoff J, LaCroix K, Yuen N, et al. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf 2006; 29 (10): 875-87
21. Lilienfeld D, Nicholas S, Macneil D, et al. Violation of homogeneity: a methodologic issue in the use of data mining tools. Drug Saf 2003; 26 (5): 363-4
22. Hauben M, Madigan D, Gerrits CM, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005; 4 (5): 929-48
23. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959 (22): 719-49
24. Breslow N. Odds ratio estimators when the data are sparse. Biometrika 1981; 68 (1): 73-84
25. Bate A. The use of a bayesian confidence propagation neural network in pharmacovigilance [doctoral thesis]. Umeå: Department of Pharmacology and Clinical Neuroscience, Umeå University, 2003
26. Bentzen N. An international glossary for general/family practice. Fam Pract 1995; 12 (3): 341-69
27. Henderson AR. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. Ann Clin Biochem 1993; 30 (Pt 6): 521-39
28. Bate A, Edwards IR, Lindquist M, et al. Violation of homogeneity: a methodological issue in the use of data mining tools. Drug Saf 2003; 26 (5): 363-6
29. Levine JG, Tonning JM, Szarfman A. Reply: the evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases: lessons to be learned. Br J Clin Pharmacol 2006; 61 (1): 105-13

30. Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: new systematic tools for an old problem. Pharmacotherapy 2004; 24 (9): 1099-104
31. Egberts AC, Meyboom RH, van Puijenbroek EP. Use of measures of disproportionality in pharmacovigilance: three Dutch examples. Drug Saf 2002; 25 (6): 453-8
32. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003; 158 (3): 280-7

Correspondence: *Johan Hopstadius*, Uppsala Monitoring Centre, Box 7051, Uppsala, S-751 40, Sweden.
E-mail: johan.hopstadius@who-umc.org