Drug Safety 2008; 31 (11): 1049-1052
0114-5916/08/0011-1049/$48.00/0

Stratification for Spontaneous Report Databases

Stephen J.W. Evans

The London School of Hygiene & Tropical Medicine, London, UK

Drug Safety has published two studies, Woo et al.[1] earlier this year and Hopstadius et al.[2] in this issue, on stratification used in screening databases of reports of suspected adverse drug reactions for 'signals'. These studies have reached somewhat different conclusions. In this commentary I will give a brief description of the issues involved and attempt to explain their different results and interpretation.

1. The Issues

In the past decade, databases of reports of individual cases where an adverse event associated with a medicine or a vaccine ('drug' is used to cover either in this article) have grown in size substantially. During this period, several different methods of detecting 'signals' of new possible adverse drug reactions have been utilized to aid in the processing of large volumes of reports. Each of these methods has used the entire database to interpret the reports for a particular drug, rather than using other data, in particular the number of prescriptions as a denominator to obtain an incidence rate. It is analogous to using only total deaths in a population to interpret deaths from a particular cause in a particular group – well known in epidemiology as a 'proportional mortality ratio'. Proportional reporting ratios (PRRs) and reporting odds ratios (RORs) relate strongly to this concept.

For each of the methods, when focusing on a particular drug/event combination under consideration as a 'signal', the observed numbers for that combination are compared with the expected numbers derived from the entire database. Larger than expected numbers are then potentially a signal – often called a 'signal of disproportionate reporting'.[3] Each reported drug/adverse event is then assessed in the same way, so a very large number of combinations is assessed and a suitable 'cut-off' criterion is used to define a signal for further evaluation. There are two major types of method: the first might be regarded as simpler 'frequentist' methods (PRRs and RORs) and the second uses Bayesian 'shrinkage', based on combining a prior probability that the vast majority of reported drug/event combinations are not adverse drug reactions (caused by the drug) but are simply coincident events. The two main methods are the Empirical Bayes Geometric Mean (EBGM) and the information component (IC) as used by the Uppsala Monitoring Centre for WHO data. Almenoff et al.[4] give a survey and further details of each of the methods.

The first, and usual, implementation of the EBGM used a stratification of the US FDA Adverse Event Reporting System (AERS) database, but the initial IC, ROR and PRR methods did not routinely use stratification. The use of stratification in epidemiology is a classic way to reduce confounding. That is, there is a variable that is associated with both the exposure (the drug) and the outcome (the adverse event). In modern epidemiology, it tends to be more usual to use logistic regression to allow for potential confounders. The natural measurement scale for the effect of the exposure is then the (log) odds ratio, analogous to the ROR. Whether stratification or regression is used in the simplest form, the assumption is that the exposure has an underlying constant effect across the different levels of confounders and that each confounder acts indepen-

dently. It is usual to test whether there is evidence of departure from such homogeneity of effect when stratifying, but this is less often done (for no good reason) when using regression. When the drug effect differs with different settings, this is called effect modification in epidemiology, which is an interaction in statistical terms. Regression methods can easily include terms to test and allow for such an interaction, but it complicates the interpretation. Analysis of data in subgroups (which effectively assumes there is effect modification) is often done wrongly in epidemiology and even in randomized trials![5]

In deciding on a 'signal', there is an arbitrary choice of cut-off. This choice will result in a particular level of false signals and of missed signals (false-positives and false-negatives in the context of diagnostic testing, or type I and type II errors in statistical testing). It is inevitable that the change in cut-off will increase one at the expense of the other. It is clear, and information theory shows, that the only way to reduce both types of error simultaneously is to add extra information. Hence, the addition of confounding variables in the analysis is one way of adding extra information. The problem of a 'gold standard' of assessment is particularly noted by Hopstadius et al.;[2] whether the Institute of Medicine list or *Reactions Weekly* are adequate is doubtful, but the issue with signals is that they are preliminary by definition. To some degree in this area, the cut-off should be decided on the basis of feasibility of further study, the resources available and the medical or public health importance of the drug/event combination under consideration. It seems likely that the problems with medicines and vaccines are different; vaccines tend to be given to otherwise healthy people, especially children. This means that the spectrum of adverse events reported will be quite different; the likelihood that any particular event is causal may be low, but tolerance for even very low levels of risk will be very low. However, the use of only vaccine data as background will mean it is more difficult to signal an event as a reaction if that event is reported commonly for other vaccines. Hence, restriction to vaccines or stratification by vaccines/medicines will have both gains and losses.

2. The Findings

Woo et al.[1] found that stratification of their data on vaccine adverse events (sourced from the US Vaccine Adverse Event Reporting System [VAERS]) has an expected effect. Of course, their data are already stratified compared with most national databases and the WHO database in that it is for vaccines only and not all medicines. They used four age groups (and unknowns) by sex as strata. The finding that many signals disappeared when stratification is used is to be expected if there is confounding. It is also not unexpected that some signals are masked by confounding but these happen less often than false-positive results. What is perhaps interesting is the differential effect with EBGM compared with PRRs. There are 283 of 457 (62%) signals removed with EBGM but 701 of 1735 (40%) with PRRs. The proportion of signals masked by confounding is also much higher with PRRs (8%) than with EBGM05 (the lower bound of the 90% CI of the EBGM) [3%]. It is possible that this may be because of what is effectively a different cut-off for PRRs than for EBGM. The PRR cut-off is approximately equivalent to having a PRR025 (the lower bound of the 95% CI of the PRR) >1 rather than an EBGM05 >2. It would be interesting to see the effect if equivalent cut-offs were used.

The approach of Hopstadius et al.,[2] using data on drug adverse events from the WHO database, has been to use unstratified analysis for their initial, signal-detecting examination of their database. Their use of the IC has been very successful in analysis of their data. It is, as they say, based on a relatively complex formula, but it is very useful to note that it is approximately a much simpler formula (their equation 1). They use a different cut-off to EBGM in that it is IC025 (the lower bound of the 95% CI of the IC). They also emphasise that their total process is a knowledge discovery process in which the signals based on the IC are not only subject to triage but may also focus attention on some drug/event combinations that are not signalled

by IC025. They note that "confounding can only be evaluated in the absence of effect modification."[2] This seems to be going too far. In epidemiology, where the precise estimate of effects is important, this might be relevant, but with signal detection this is not necessary. It would be very unlikely that a drug would have an effect of causing an ADR in one stratum but be protective against that effect in another stratum. If the effect is in the same direction but of varying magnitude it does not really matter provided the signal is detected. Perhaps all signalling methods should incorporate a test for homogeneity of effect across strata, with a warning to suggest further study when there is heterogeneity.

The general argument that having too many strata again seems to be reasonable, although that does not necessarily mean that stratification should not be carried out. The use of reporting quarter as well as age, sex and country seems likely to lead to too many strata; use of reporting year, or even say 5-year groups, would reduce the numbers. This over-stratification may have contributed to the decreased sensitivity found by Hopstadius in his thesis.[6]

The choice of potential confounders in ADR surveillance cannot be wide, since for many reports these will not be recorded, and even co-medication, which is very important, is often missing. However, it does seem that, if this could be taken into account using shrinkage regression methods as noted by Hopstadius et al.,[2] then there would be gains since adding extra information could improve both sensitivity and specificity.

The simulation study provides results that are not unexpected. Firstly, having 34 756 strata is just not sensible. Secondly, random allocation of reports that ignores their actual source in terms of stratum simply generates noise. Perhaps I have misunderstood, but the suggestion it "should have no true impact on the observed-to-expected ratio"[2] does not seem to be correct. It seems very likely that it will bias any effects towards a null effect.

The study of the impact of potential confounders uses eight age categories plus unspecified; this is double the number used in the study on VAERS, and is actually larger than the number used in the earlier simulation. Perhaps age beyond 45 years has little confounding effect in ADR reporting. Combining countries with small amounts of reporting so ensuring that each stratum has >100 reports seems sensible.

A further problem seems to be with the idea that "any small strata"[2] cause a problem. The correct use of weights should ensure that a small weight is given to such small strata. Propensity scores have few advantages over regression adjustment in general but the problems noted by Hopstadius et al.[2] about computing these separately for each drug show that the methods are not likely to be useful.

3. Conclusions

Both of the papers considered here are used by public bodies where many products are being used, in contrast to company-specific databases. It is not yet clear whether the conclusions from either group can be applied to smaller databases.

Stratification can add useful information in detecting signals. Small strata should be avoided and there needs to be concentration on those that show strong effects when applied to a particular dataset.

Simplicity and transparency have some advantages, but the process of dealing with analysis of spontaneous reports requires more than just a single unthinking application of algorithms. The whole system used by WHO provides a good lesson.

Acknowledgements

No sources of funding were used to assist in the preparation of this commentary. The author has no conflicts of interest that are directly relevant to the content of this commentary.

References

1. 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
2. Hopstadius J, Norén GN, Bate A, et al. Impact of stratification in adverse drug reaction surveillance. Drug Saf 2008; 31 (11): 1035-48
3. Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms [published erratum appears in Drug Saf 2006; 29 (12): 1192]. Drug Saf 2004; 27 (10): 735-44

4. Almenoff JS, Pattishall EN, Gibbs TG, et al. Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharmacol Ther 2007; 82 (2): 157-66
5. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine: reporting of subgroup analyses in clinical trials. N Engl J Med 2007; 357 (21): 2189-94
6. 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

Correspondence: Professor *Stephen J.W. Evans*, The London School of Hygiene & Tropical Medicine, Keppel St, London, WC1E 7HT, UK.