Drug Safety 2008; 31 (12): 1145-1147
0114-5916/08/0012-1145/$48.00/0

Stratification for Spontaneous Report Databases

We thank Evans[1] for the insightful and valuable comments on our paper concerning the use of routine adjustment by stratification in adverse drug reaction surveillance.[2] Both our paper and that of Woo et al.[3] consider the implications of routine adjustment by stratification in first pass screening of large collections of individual case safety reports (ICSRs) and find that routine adjustment by stratification tends to lower the number of drug-ADR pairs highlighted as excessively reported together. Woo et al.[3] assumed that discrepancies between the adjusted and crude measures were solely attributable to eliminated confounding, and that drug-ADR pairs highlighted exclusively by the crude analysis were indeed false-positives. In our paper,[2] we evaluated the differences due to adjustment relative to a gold standard, defined for the study as first published case reports of suspected ADRs in the international medical literature, as listed in *Reactions Weekly* (Wolters Kluwer Health | Adis), and found that stratification only increased the efficiency marginally. Similarly, Jakobsson[4] analysed the effects of stratification on vaccines in the WHO ICSR database, VigiBase, and found several examples where large discrepancies between adjusted and crude measures of association did not simply correspond to successful confounder elimination. While we concur with Evans[1] that in pharmacovigilance such extreme examples are unlikely to be due to situations where the clinical effects point in different directions for different subgroups, we disagree that they are unworthy of further attention; since they may well represent important reporting idiosyncrasies. For example, in Jakobsson's study[4] the unadjusted information component (IC) between influenza vaccine and aggravated rheumatoid arthritis was 1.8 but the IC adjusted for age, country and time of reporting (1843 strata) was −0.6 (see table I). According to the logic employed by Woo et al.,[3] this would be considered as an example of a false-positive due to confounding, eliminated through adjustment by stratification. However, it is clear from the underlying data for this example, reproduced in table I, that except in a small portion of the database, the drug and the ADR are reported together unexpectedly often. A very large part of the total expected count of 22.81 in the adjusted analysis comes from one specific country and time period (Canada, 2000–4). This emphasizes the potential vulnerability of adjustment by stratification to data quality issues. In this example, routine adjustment by stratification would have hidden a potentially important reporting pattern present in a considerable part of the database.

Our paper[2] shows that the adjusted observed-to-expected ratio, as implemented for routine use in both the Empirical Bayes Geometric Mean (EBGM) and the IC[5,6] is vulnerable to the presence of any very small strata. With this type of adjustment, the stratum specific observed-to-expected ratios are weighted by the stratum specific expected numbers of events. The problem with this weighting is that it may attribute too high weights to very small strata for which the computed expected numbers of events are not reliable. We agree with Evans[1] that this undesirable property is not a prerequisite for the calculation of an adjusted measure. Indeed, Gould[7] proposed an alternative weighting by the size of each stratum. The disadvantage of such a weighting is that it may give high weights to unreliable stratum specific observed-to-expected ratios from large strata in which either the drug or the ADR do not occur or are unusually rare.[6] Overall, we believe the weighting by the expected number of events is to be preferred, on the condition that small strata are avoided in the stratification.

In our paper[2] we have shown that there is a theoretical basis for an over-stratification leading to decreased sensitivity, and our simulation was intended to demonstrate this. Figures 1a and 1b, rescaled from our paper,[2] show simulations based on random reallocation of all reports in the WHO database such that the same numbers of reports in each predefined stratum is identical to that in the real dataset: with granular stratification (figure 1a) and with coarser stratification (figure 1b). Whatever the

Table I. Stratum specific data for influenza vaccine and rheumatoid arthritis aggravated. All strata where either the observed or the expected number of reports is larger than 0.5 are listed

Age, time of reporting and country of origin	Observed number of reports	Expected number of reports	IC	IC025
45–4, 1997–9, GBR	2	0.05	2.19	−0.30
75–, 1991–3, GBR	2	0.04	2.20	−0.26
45–64, 1994–6, GBR	1	0.09	1.36	−2.35
65–74, 2003–4, GBR	1	0.07	1.41	−2.29
65–74, 1994–6, GBR	1	0.06	1.42	−2.27
75–, 1997–9, NZL	1	0.04	1.47	−2.14
65–74, 1997–9, GBR	1	0.03	1.51	−2.20
25–44, 2000–2, AUS	1	0.02	1.53	−2.18
5–14, 1991–6, NLD	1	0.01	1.55	−2.13
75–, 1994–6, SWE	1	0.01	1.55	−2.15
45–64, 2000–4, NLD	1	0.01	1.57	−2.16
Unknown, 2005–7, NLD	1	0.01	1.57	−2.16
25–44, 1997–9, AUS	1	0.00	1.57	−2.15
45–64, 2003–4, CAN	0	10.65	−4.48	−15.12
25–4, 2003–4, CAN	0	2.45	−2.56	−13.18
65–74, 2003–4, CAN	0	2.15	−2.41	−13.03
45–64, 2000–2, USA	0	1.77	−2.18	−12.84
Unknown, 2003–4, CAN	0	0.90	−1.49	−12.12
65–74, 2000–2, USA	0	0.79	−1.37	−12.02
75–, 2003–4, CAN	0	0.64	−1.19	−11.79
25–44, 2000–2, USA	0	0.52	−1.02	−11.67

AUS = Australia; **CAN** = Canada; **GBR** = Great Britain; **IC** = information component; **IC025** = the lower bound of the 95% CI of the IC; **NLD** = the Netherlands; **NZL** = New Zealand.

underlying causes are for the magnitude of effect, it is clear that high levels of stratification lead to less sensitivity in signal detection.

Our simulation study[2] was based on a 'naive' stratification with 34 000 strata including many very small ones. While Evans[1] commented that this seemed a large number of strata to consider, we note that Szarfman and colleagues[8,9] have suggested that the routine implementation of EBGM in Multi-item Gamma Poisson Shrinker (MGPS) on US FDA data should use more than 1000 strata and that such stratification is a strength of MGPS.[10,11] Previous work on VigiBase data has suggested that country is one of the most influential sources of confounding.[12-14] A combination of the standard stratification used by the FDA with stratification by the 89 different reporting countries would lead to more than 20 000 strata in VigiBase. That said, a fundamental message of our paper is that the total number of strata should be kept low and the size of the smallest strata high. Our real-world analysis achieved this by simultaneous adjustment by country of origin and time of reporting (see figure 1b) and gave a useful result for our purposes.

We concur with Evans' argument from information theory that stratification should allow for improved screening capability, and that confounder detection should be a key component in the overall knowledge discovery process in adverse drug reaction surveillance. Using measures adjusted by stratification can certainly provide practical benefits. For example, Jakobsson[4] gave several examples for live oral polio vaccine where adjustment for patient age provided a more appropriate overall measure of association. However, it is our view that optimal first-pass screening can only be achieved in a transparent knowledge discovery framework using a combination of crude, adjusted, and stratum specific measures of association. Adjusted measures of association must be transparent with regard to the choice

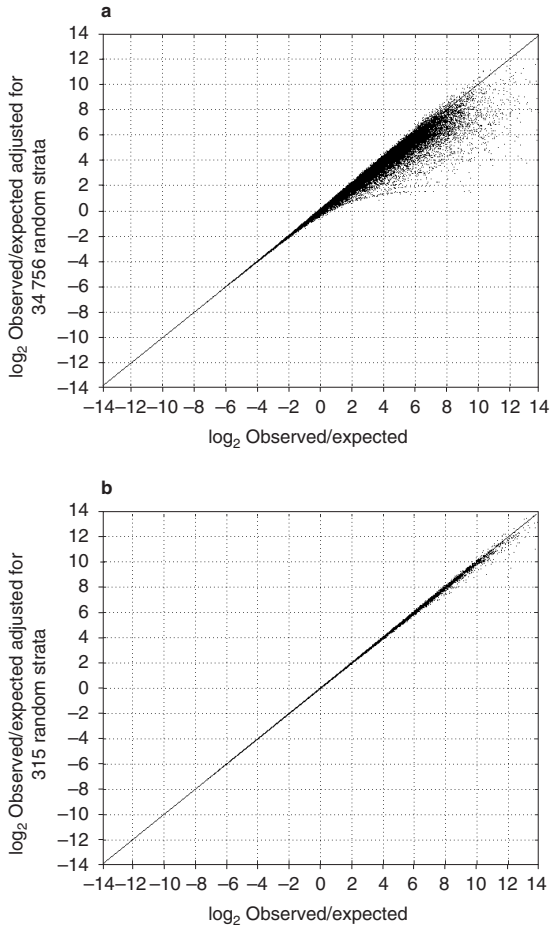

Fig. 1. Adjusted vs crude observed-to-expected ratios for a random allocation of all reports to strata, such that the number of reports in each stratum was kept the same as when stratifying Vigibase on (**a**) patient sex, patient age (eight groups), country of origin and time of reporting (in quarterly intervals) using 34 756 strata in total; and (**b**) country of origin and time of reporting (nine groups) using 315 strata in total. To avoid small strata when combining time of reporting and country of origin, countries with less than 100 reports were merged and time periods were automatically merged for countries with little or irregular reporting until each country-time stratum contained at least 100 reports.

of stratification variables and their intervals as well as to which strata have been particularly influential on the overall adjusted measure. This will allow for a distinction between successful confounder elimination and inappropriate masking of stratum specific effects.

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

Acknowledgements

The authors have no conflicts of interest that are directly relevant to the content of this letter.

References

1. Evans SJ. Stratification for spontaneous report databases. Drug Saf 2008; 31 (11): 1049-52
2. Hopstadius J, Norén GN, Bate A, et al. Impact of stratification on adverse drug reaction surveillance. Drug Saf 2008; 31 (11): 1035-48
3. 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
4. Jakobsson M. Vaccine pharmacovigilance in the WHO database [masters thesis]. Uppsala: Uppsala Universitet, 2008
5. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Statistician 1999; 53 (3): 177-90
6. 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
7. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf 2003; 12 (7): 559-74
8. Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: New systematic tools for an old problem. Pharmacotherapy 2004; 24 (9): 1099-104
9. 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
10. DuMouchel W, Smith ET, Beasley R, et al. Association of asthma therapy and Churg-Strauss syndrome: an analysis of postmarketing surveillance data. Clin Ther 2004; 26 (7): 1092-104
11. Hauben M, Reich L. Response to letter by Levine, et al. Br J Clin Pharmacol 2006; 61 (1): 115-7
12. 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
13. Bate A, Edwards I, Lindquist M, et al. Violation of homogeneity: a methodological issue in the use of data mining tools [letter]. Drug Saf 2003; 26 (5): 363-6
14. Hopstadius J. Methods to control for confounding variables in screening for association in the WHO drug safety database [masters thesis]. Uppsala: Uppsala University, 2006