Drug Safety 2003; 26 (5): 363-364
0114-5916/03/0005-0363/$30.00/0

Violation of Homogeneity: A Methodologic Issue in the Use of Data Mining Tools

The recent publication by Bate et al.[1] suggests that use of the 'method Bayesian confidence propagation neural network (BCPNN)' facilitates the identification of signals in the WHO database. As described by Lindquist et al.,[2] the technique is premised on the appropriate use of proportionate reporting rates. Such use makes an implicit assumption regarding the homogeneity of the population to which it is applied. Bate et al.[1] describe the WHO database as one in which 67 countries provide data. We suggest that the assumption of homogeneity is violated by the use of data from so many countries representing diverse medical care delivery systems, regulatory environments, pharmaceutical utilisation, and pharmacovigilance systems. It seems reasonable to conclude that use of BCPNN may result in the identification of potential signals but also create considerable noise.

The problems presented by the use of heterogeneous populations in situations similar to the present one were first explored by Cochran.[3] Subsequent work on the problem by Mantel and Haenszel[4] resulted in many methodological advances.[5] The epidemiological community has been aware of the challenges presented by the use of summary measures generated from heterogeneous data (usually referred to as either 'crude' or 'unadjusted') for more than a century[6] and since the innovations of both Cochran,[3] and Mantel and Haenszel[4] has not relied upon such an approach. It is standard practice in pharmacoepidemiological studies to adjust such data to account for the heterogeneity. If the BCPNN is to be used to advantage in the identification of signals from the WHO database, it will be essential to develop some means of adjusting the technique to account for the heterogeneity present in those data. The alternative is that much effort will be expended investigating noise with little consequent protection or improvement in the public's health. This alternative is not desirable, as it undercuts both the scientific and necessary social and political support for pharmacovigilance activities.

David E. Lilienfeld
Bristol-Myers Squibb Company, Pharmaceutical
Research Institute, Global Epidemiology and
Outcomes Research, Princeton, New Jersey, USA

Savian Nicholas
Bristol-Myers Squibb Company, Pharmaceutical
Research Institute, Global Pharmacovigilance and
Labelling, Princeton, New Jersey, USA

Daniel J. Macneil
Bristol-Myers Squibb Company, Pharmaceutical
Research Institute, Global Pharmacovigilance and
Labelling, Princeton, New Jersey, USA

Olga Kurjatkin
Bristol-Myers Squibb Company, Pharmaceutical
Research Institute, Global Pharmacovigilance and
Labelling, Princeton, New Jersey, USA

Thomas Gelardin
Bristol-Myers Squibb Company, Pharmaceutical
Research Institute, Global Pharmacovigilance and
Labelling, Princeton, New Jersey, USA

References

1. Bate A, Lindquist M, Edwards IR, et al. A data mining approach for signal detection and analysis. Drug Saf 2002; 25 (6): 393-7
2. 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
3. Cochran WG. Some methods for strengthening the common X2 tests. Biometrics 1954; 10: 417-51
4. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959; 22: 719-48
5. Fienberg SE. The analysis of cross-classified categorical data. 2nd ed. Cambridge (MA): MIT Press, 1980
6. Lilienfeld DE. The greening of epidemiology: sanitary physicians and the London Epidemiological Society. Bull Hist Med 1979; 52: 503-28