Drug Saf (2015) 38:517–518
DOI 10.1007/s40264-015-0293-9

EDITORIAL

Is the Yellow Card Road Going in the Right Direction?

Stephen J. W. Evans[1]

Published online: 7 May 2015

In 1963, following the dreadful consequences of giving thalidomide to pregnant women, the UK Government set up the Committee on Safety of Drugs, chaired by Sir Derrick Dunlop. In May 1964, he wrote to all UK doctors (a mammoth task with no information technology to support it) asking them to report "any untoward condition in a patient that *might* [his italics] be the result of drug treatment". He also promised that such reports "will be treated with complete professional confidence" and "will never be used for disciplinary purposes or for enquiries about prescribing costs". The UK Medicines and Healthcare products Regulatory Agency has recently 'celebrated' the 50th anniversary of what became known as the 'Yellow Card System' for reporting suspected adverse reactions to medicines. Professor David Finney (born January 1917) was at that 'celebration' and he was a pioneer in statistical approaches to assessing the reports, with his first paper in the area being published in 1963. He also worked with the WHO, where, in 1969, Patwary [1] wrote an internal report (confidential at the time) that set out many of the statistical principles in utilising these 'spontaneous reporting' data, and these were clearly summarised by Finney [2]. The ability to analyse all the reports was limited by the information technology of the time, but more than 40 years on, the situation has changed dramatically. The strictures on confidentiality may have prevented independent statisticians applying their minds to make best use of the data but, worldwide, there is now much greater openness.

Most spontaneous reports are now entered into electronic databases and many countries and most large pharmaceutical companies apply statistical methods to detect possible signals of adverse drug reactions (ADRs). It is widely acknowledged that signals produced from an automated system will need further assessment before being more widely propagated as a 'signal' in the usual sense used in pharmacovigilance; therefore, a signal detected solely by using statistical methods is usually referred to as a 'signal of disproportionate reporting' (SDR) [3, 4]. Many also apply the term to vaccines and apply similar methods to databases that contain reports of suspected adverse reactions to vaccines. There are a number of different statistical methods applied to this type of database, and there have been a number of evaluations of the methods, usually applied to a single database of spontaneous reports.

The paper by Candore et al. [5] is an evaluation of five methods, different, to some degree, in principle, with multiple cut-off criteria to define a 'signal', in seven different databases. The databases included specific companies, a specific country, and two international databases. They are not totally independent sets of data since some reports will appear in multiple databases. They have focussed on a subset of the possible SDRs and, in particular, have chosen to examine a limited set of products, with 220 the maximum studied, but for which only two databases had a complete set. Inevitably, the company databases are limited to products marketed by the particular companies.

The SDRs produced by the different methods have been evaluated as to whether they are a true positive or a false positive, based on whether there is acceptance of the adverse event being a true ADR by being included in product

✉ Stephen J. W. Evans
stephen.evans@Lshtm.ac.uk

[1] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1 7HT, UK

information for the drug. Of course this is not an absolute standard, since knowledge of the truth is inevitably limited, but is probably the best that can be done. It does mean there are some limitations; therefore, some events will be classified as ADRs when they are not truly caused by the drug, and some, not classified as ADRs, may actually be caused by the drug but have not been accepted as causal. The exclusion of products that have recently been marketed minimises this latter risk.

The signals are assessed overall by the positive predictive value (here called the 'precision'), which is obviously maximised by few false positives. The failure to detect a signal for ADRs included in product information is measured by the sensitivity—the proportion of known ADRs for which an SDR was reported at some stage. The authors use these measures to reflect what is important to those detecting signals for which there are two questions. First, does the method find things that are, or turn out to be, genuine ADRs (precision)? Second, does it do this without producing many signals that are not true ADRs (sensitivity)?

The overall results make it clear that these questions are answered somewhat positively; however, what determines success is less the actual statistical method but rather the choice of a cut-off point. Both the Bayesian methods (and shrinkage applied to the reporting odds ratio) tested here shrink the strength of a signal based on small numbers of observations towards a null (no signal) value, while the frequentist methods have cut-off values that take small observed numbers into account, both directly by not counting as a signal disproportionate values based on small numbers and indirectly through the use of a confidence interval or the essentially equivalent Chi-squared value. Figure 6 in the study by Candore et al. [5] clearly shows that varying the cut-off criteria can result in almost any value of precision and sensitivity that can be obtained by any method but using just a single method (proportional reporting ratio in this case). There will be a penalty in some instances in terms of having a longer delay in first detecting an SDR with higher thresholds, but the converse is true in that varying the criteria can obtain earlier detection if desired, as shown in Fig. 7 of the study by Candore et al. [5].

The main lessons are that (1) the choice of a method cannot be dictated universally but should be tailored to a particular database and the use that is to be made of the signals that are detected; and (2) the performance of all the methods appears to decline as a product continues on the market for a longer time. The first lesson is not very surprising and has been suspected for some time. Each of the methods, as the authors point out, is based on the same data—the observed and expected being based on a simple 2×2 table. A method that uses differences rather than ratios is available [6], but evaluation of that method has not shown any particular advantage [7].

Where Do We Go from Here?

There is a possibility that new methods for use in spontaneous reporting databases might result in improved properties, but the gains are likely to be marginal. Information theory tells us that the only way of improving false positives and false negatives simultaneously is to add extra information. Whether adding extra data does improve things, at least in terms of using stratification, still seems controversial.

It is not clear that the hopes we had that electronic health record databases would be a great deal better may not be realised (see the special issue of *Drug Safety* regarding the Observational Medical Outcomes Partnership [OMOP] project [8]). Use of both spontaneous reporting and electronic health record databases may be the best way forward. Allowing for a high false positive rate in spontaneous reporting signals may not be a problem if they can be rapidly checked in electronic health record databases. The problem may be that the electronic health record databases may not have enough exposure to new drugs, and the potential 100 % coverage with spontaneous reports may mean we will need to go on relying on them for some time yet. We have not reached the end of the Yellow Card road, and it seems the direction is the right one at the moment!

Compliance with ethical standards No sources of funding were used to assist in the preparation of this editorial. Stephen Evans has no conflicts of interest that are directly relevant to the content of this editorial.

References

1. Patwary KW. Report on statistical aspects of the pilot research project for international drug monitoring. Geneva: WHO; 1969.
2. Finney DJ. Systematic signalling of adverse reactions to drugs. Methods Inf Med. 1974;13:1–10.
3. Hauben M, Reich L, Chung S. Postmarketing surveillance of potentially fatal reactions to oncology drugs: potential utility of two signal-detection algorithms. Eur J Clin Pharmacol. 2004;60:747–50.
4. Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf. 2009;32:99–110.
5. Candore G, Juhlin K, Manlik K, et al. Comparison of statistical signal detection methods within and across spontaneous reporting databases. Drug Saf. doi:10.1007/s40264-015-0289-5. (In press).
6. Evans SJW, Nitsch D. Statistics: analysis and presentation of safety data. In: Talbot J, Aronson JK, editors. Stephens' detection and evaluation of adverse drug reactions: principles and practice. 6th ed. Chichester; Wiley; 2011. p. 371–373.
7. 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.
8. Evans SJ. Moving along the yellow brick (card) road? Drug Saf. 2013;36(Suppl 1):3–4.