ORIGINAL RESEARCH ARTICLE

Drug Safety 2005; 28 (10): 901-906
0114-5916/05/0010-0901/$34.95/0

Testing and Implementing Signal Impact Analysis in a Regulatory Setting

Results of a Pilot Study

Emma Heeley,[1] *Patrick Waller*[2] and *Jane Moseley*[3]

1 Post-Licensing Division, Medicines and Healthcare products Regulatory Agency, London, UK
2 Patrick Waller Limited, Consultancy in Pharmacovigilance and Pharmacoepidemiology, Southampton, UK
3 Post-Licensing Division, Pharmacoepidemiology Research Team, Medicines and Healthcare products Regulatory Agency, London, UK

Abstract

Background and aim: Statistical signal detection methods such as proportional reporting ratios (PRRs) detect many drug safety signals when applied to databases of spontaneous suspected adverse drug reactions (ADRs). Impact analysis is a tool that was developed as an aid to prioritisation of such signals. This paper describes a pilot project whereby impact analysis was simultaneously introduced into practice in a regulatory setting and tested in comparison with the existing approach.

Methods: Impact analysis was run on signals detected during a 26-week period from the UK Adverse Drug Reactions On-line Information Tracking (ADROIT) database of spontaneous ADRs that met minimum criteria (PRR ≥3.0, χ^2 ≥4.0 and ≥3 reported cases) and related to established drugs (i.e. those that have been available for at least 2 years and no longer carry the 'black triangle' symbol). The current method of signal prioritisation (i.e. the collective judgement at a weekly meeting) was initially performed without knowledge of the findings of impact analysis. Subsequently, the meeting was presented with the findings and, where appropriate, given the opportunity to reconsider the judgement made. The categories arising from the two methods were compared and the ultimate action recorded. Inter-observer variation between scientists performing impact analysis was also assessed.

Results: Eighty-six separate signals were analysed by impact analysis, of which 5% were categorised as high priority (A), 14% as requiring further information (B), 31% as low priority (C) and 50% as no action required (D). In general, the new method tended to give a higher level of priority to signals than the existing approach. Overall, there was 59% agreement between the impact analysis and the collective judgement at the meetings (kappa statistic = 0.30). There was slightly greater agreement between impact analysis and the final action taken (kappa statistic = 0.39), indicating that the findings of an impact analysis had an influence on the outcome. Assessment of inter-observer variation demonstrated that the method is repeatable (kappa statistic for overall category = 0.77). Almost 70% of those who participated in the pilot study believed that impact analysis represented an improvement in how signals were prioritised.

Conclusions: Impact analysis is a repeatable method of signal prioritisation that tended to give a higher level of priority to signals than the standard approach and which had an influence on the ultimate outcome.

Background

Impact analysis is a tool that was developed to prioritise signals arising from spontaneous adverse drug reaction (ADR) reporting data and is described by Waller et al.[1] In this paper we describe a pilot study conducted at the UK Medicines and Healthcare products Regulatory Agency (MHRA). The objectives of the study were to test the feasibility and reproducibility of impact analysis and to compare it with the existing approach to signal prioritisation (a collective judgement made at review meetings).

Each week the MHRA receives >300 spontaneous ADR reports on 'yellow cards' from UK health professionals and the pharmaceutical industry. The main purpose of the yellow card scheme is to provide 'early warnings' of previously unsuspected ADRs (signals). Over the past decade various mathematical tools have been developed and applied to ADR databases to help facilitate in the identification of such signals.[2-4] The MHRA currently uses proportional reporting ratios (PRRs).[5] Such methods detect large numbers of signals and consequently there is a need to prioritise them. The methods available to prioritise or triage these signals have been based on principles[6-8] and qualitative criteria such as 'SNIP'.[9] SNIP has been used at the MHRA and takes into account the strength of the signal, whether it is really new, the clinical importance of the ADR and its potential for prevention.

The new method described in the accompanying paper prioritises signals based on their strength of evidence for causation and public health implications.[1] The output of impact analysis is to prioritise signals into one of four categories, with each implying a consequential course of action as follows: A – high priority; B – there is need to gather more information; C – low priority but still needs to be addressed; and D – no action is warranted at the present time.

In the MHRA, normal practice is that all new yellow card reports for drug substances that are not intensively monitored (i.e. are not part of the 'black triangle' scheme that applies to new drugs in the first 2 years of marketing) are screened on a weekly basis. Signals are those drug substance-ADR combinations above quantitative threshold criteria (as defined in the methods section) or where, irrespective of quantitative criteria, reports are received with a fatal outcome for reactions occurring in children or with drug interactions. These signals are subjected to preliminary analysis by pharmacovigilance scientists before discussion with experienced assessors at a weekly prioritisation meeting. An action plan for each signal is agreed based on the collective judgement of the meeting.

Methods

In this study, impact analysis was performed by pharmacovigilance scientists on signals detected in the week before the prioritisation meeting. The results of the impact analysis were compared with the collective judgement at the meeting prior to knowledge of the impact analysis result and the subsequent actions taken. Inter-observer variation between scientists performing impact analysis on the same signals was also assessed.

Pharmacovigilance scientists were trained in the use of impact analysis and provided with standard operating procedures and a user guide. Each week from 4 May to 7 November 2003, prior to the weekly signal review meeting, pharmacovigilance scientists performed impact analysis on up to two UK signals that were not labelled in the summary of product characteristics (SPC). For the purpose of this study, a signal was defined as a drug-ADR combination with a PRR ≥ 3, $\chi^2 \geq 4$ and ≥ 3 reports. At the meeting, scientists first presented their signals in the normal manner without revealing the impact analysis score. The collective judgement of the meeting was then used to define one of the four courses of action that equate to the impact analysis categories A–D (as previously described). The result of the impact analysis was then presented and the

final course of action was agreed upon. Each signal was, therefore, assigned three separate categories derived from: (i) impact analysis; (ii) initial collective judgement (prior to knowledge of impact analysis result); and (iii) the final action taken.

For assessment of inter-observer variation, a sample of drug event combinations was analysed. Pharmacovigilance co-ordinators (experienced pharmacovigilance scientists) were asked to, independently and without knowledge of the result, duplicate the assessment of the two inputs that require a judgement to be made, i.e. the strength of the cases and biological plausibility. The impact analysis category was then recalculated using these values.

Kappa statistics were used to assess the levels of agreement between the impact analysis category, collective judgement and the final action taken, and to assess the degree of inter-observer variation. The results from the kappa analysis were categorised into three groups based on the strength of agreement proposed by Landis and Koch,[10] as shown in table I.

Table I. Categories of the strength of agreement for the kappa analysis based on those proposed by Landis and Koch[10]

Result for kappa	Strength of agreement
<0.00–0.30	Poor
0.31–0.60	Fair
0.61–1.00	Good

Results

The pilot study ran for 26 weeks and impact analysis was performed on 87 signals. Fourteen pharmacovigilance scientists performed impact analysis, each completing a mean of 6.2 analyses (median 5.5, range 1–20). The 87 signals related to 59 drug substances and 72 different ADRs (as defined at Adverse Drug Reactions On-line Information Tracking [ADROIT] 'preferred term' level). There were 86 drug ADR combinations; one drug-ADR combination was rerun during the study period because another report had been received.

Distribution of Impact Analysis Categories

The distribution of the categories of the 87 signals analysed by impact analysis is shown in table II. Only 4 of the 87 signals analysed were categorised as A (high priority), whereas 50% of the signals were categorised as D (no further action warranted at this time). Figure 1 illustrates the distribution of the 87 signals in respect of scores for evidence and public health. The four quadrants of the graph represent the four categories of impact analysis. Those signals with an evidence score of >10 and a public health score of >10 are in the top-right-hand quadrant and categorised as A. It is possible that a signal can have a public health score or an evidence score from 1 to 100. In this study there was a lesser range of public health scores (1–18) than evidence scores (1–60).

Table II. Distribution of the impact analysis categories for the 87 signals analysed

Category	Impact analysis [n (%)]
A	4 (4.6)
B	12 (13.8)
C	27 (31.0)
D	44 (50.6)
Total	**87**

Method Comparison

The method comparison study was carried out on 86 signals as one signal was not assigned a collective judgement. For 51 of the 86 signals (59%) there was agreement between the impact analysis and the collective judgement at the meeting (kappa statistic 0.30) [figure 2]. There was slightly greater agreement between impact analysis and the final action taken (65% agreement, kappa statistic 0.39).

Discrepancies Between Impact Analysis and the Collective Judgement

There were 35 of the 86 signals where the impact analysis result did not agree with the collective judgement. In 22 cases they differed by one category, in 13 cases by 2 categories and in no case by 3 categories (i.e. no signal was categorised as an A by one method and as a D by the other). Overall, impact analysis tended to rank the signals with a higher prioritisation than the collective judgement of the panel, with 25 signals graded lower and 10 signals graded higher by the collective judgement compared with the impact analysis. For the 10 (12%) signals graded higher by the collective judgement, the final action taken for three of these signals agreed with

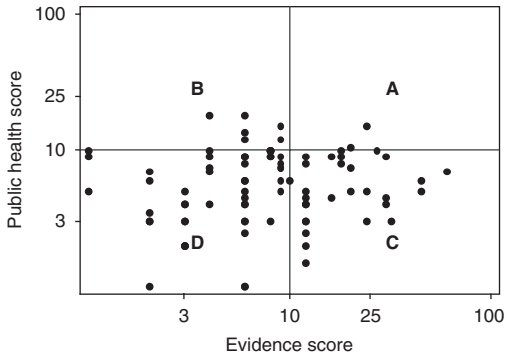

Fig. 1. Evidence and public health scores plotted for each of the 87 signals analysed by impact analysis. The cut-off values of ten are illustrated by lines and the impact analysis categories are indicated in each of the quadrants.

the impact analysis result. For a further three signals no decision on a category for final action has been made yet, as more data are expected.

There were two signals categorised as a D by impact analysis and a B by the collective judgement. For both of these signals, the meeting decided that although that particular preferred term was not a strong enough signal to warrant action, combined with the evidence for related terms (i.e. the high-level term) there was enough evidence. These preferred terms were 'systemic lupus erythematosus rash' and 'suicidal ideation'.

In the final action taken, the collective judgement was not swayed by the impact analysis result for any of the 25 (29%) signals where the collective judgement categorised the signal as lower than the impact analysis. There were many reasons for the lower category of the collective judgement and these included the reaction being strongly confounded, other cosuspect drugs and there being similar reactions already labelled in the SPC.

Sensitivity Analysis

The purpose of the sensitivity analysis is to allow for uncertainty in the data (especially biological plausibility and the strength of cases that require a subjective judgement to be made) and to allow users to gauge the possible impact of a change in input value on the results. Therefore, the sensitivity analysis was used to further look at those signals where the impact analysis result disagreed with the collective judgement and to identify if the collective judgement category was included within the findings of the sensitivity analysis. For the ten signals where the collective judgement was higher than the impact analysis results, four were included in the sensitivity analysis and six were not. Where the collective judgement was lower than the impact analysis (n = 25), the collective judgement was included in the sensitivity analysis for 16 signals but not in 9 signals. Overall, for those signals where the impact analysis disagreed with the collective judgement, 57% (20 of 35) were included in the sensitivity analysis.

Inter-Observer Variation

There were three separate analyses for the inter-observer assessment: (i) biological plausibility; (ii) case assessment; and (iii) overall impact analysis categories. Twelve scientists participated in the inter-observer variation study, four of whom acted as co-ordinators who assessed the biological plausibility and the strength of the cases independently. These two variables require a judgement to be made. Forty-eight signals were analysed by two observers (scientist and co-ordinator) for the inter-observer study.

Biological Plausibility

For 27 of 48 signals analysed (56%), there was agreement between the two observers on the biological plausibility (kappa statistic 0.33). These results are shown in figure 3. There was only one signal where the co-ordinator and scientist differed by two categories.

Strength of Cases

In the assessment of cases, the two observers agreed for 22 of 48 signals (46%) analysed (kappa statistic 0.20). The results are shown in figure 4.

		Collective judgement category				
		A	B	C	D	Total
Impact analysis category	A	1	0	3	0	4
	B	0	0	4	8	12
	C	0	3	14	10	27
	D	0	2	5	36	43
	Total	1	5	26	54	86

Fig. 2. Comparison between the categories assigned to each signal by the collective judgement of the assessors and the scientists using impact analysis.

		Co-ordinator				
		0	1	2	3	Total
Scientist	0	8	3	0	0	11
	1	5	13	3	0	21
	2	1	8	6	0	15
	3	0	0	1	0	1
	Total	14	24	10	0	48

Fig. 3. Assessment of biological plausibility by scientists and co-ordinators.

There were four signals where the assessments were different by two levels, in each case the co-ordinator graded the cases lower than the scientist.

Overall Impact Analysis Category

The observers agreed on the overall impact analysis category for 85% (41 of 48) of the signals analysed (see figure 5). This is good agreement (kappa statistic 0.77) and illustrates that although there is inter-observer variation on the subjective inputs, biological plausibility and strength of the cases, it makes little difference to the overall outcome.

Questionnaire

The scientists and those more experienced staff who formed the collective judgement where asked their view on impact analysis (response rate 81% of the 27 participants in the pilot study). Overall, the feedback was very positive, with 68% of the questionnaire responders indicating that impact analysis improved how signals are prioritised.

Discussion

This pilot study that investigated the use of impact analysis at the MHRA has demonstrated that, in this setting, it provides a repeatable, systematic and formalised approach to prioritising safety signals that arise from spontaneous ADR data. The sample size of 87 signals was sufficient to ensure that it can be applied to almost any signal arising from such data and no major problems were encountered in its application.

In theory, the method seems preferable to judgements based entirely on qualitative criteria (such as SNIP[9]). One difficulty in assessing the value of impact analysis is the absence of a gold standard for the prioritisation of signals. This pilot study showed fairly poor agreement between impact analysis and the collective judgement. However, in this study, impact analysis generally ranked signals as a higher priority than the collective judgement of a weekly signal review meeting. There are many possible reasons for this, such as signals that were selected were not new signals as action was already ongoing or assessors applied clinical judgement to the existing terminology in the SPC and decided that the specific preferred term was sufficiently labelled in the SPC. The agreement between impact analysis and the final action is better than between impact analysis and the collective judgement, representing a combination of knowledge and data from both methods. The pilot study also demonstrates that most signals (PRR ≥ 3, $\chi^2 \geq 4$, ≥ 3 reports and not labelled in the SPC) analysed in this pilot study were of low priority when assessed by either method.

The output produced from the impact analysis program not only produces a suggested consequential action but also provides a sensitivity analysis of what the action would be if each score was classified one category higher or lower than that given. This extra analysis helps the user see what would happen in the case that they were unsure of which of two categories to select, and gives confidence in the stability of the categorisation.

Impact analysis is a method for determining which signals are the most important. Its strengths lie in promoting a systematic scientific approach and in providing an accessible audit trail for the decision making process. This facility is particularly useful when signals are reassessed because further data have become available. However, there are potential drawbacks with using impact analysis. The numeri-

		Co-ordinator				
		Weak	Fairly weak	Average	Fairly strong	Total
Scientist	Weak	2	2	0	0	4
	Fairly weak	5	7	2	0	14
	Average	3	5	13	3	24
	Fairly strong	0	1	5	0	6
	Total	10	15	20	3	48

Fig. 4. Assessment of the strength of the cases by scientists and co-ordinators.

		Co-ordinator				
		A	B	C	D	Total
Scientist	A	2	0	0	0	2
	B	1	6	1	0	8
	C	0	0	10	4	14
	D	0	0	1	23	24
	Total	3	6	12	27	48

Fig. 5. Overall impact analysis category as assessed by scientists and co-ordinators.

cal values of the evidence and public health score provide a category that is only a guide. Therefore, it is important that these numerical values are not over interpreted. Furthermore, the level of the dictionary at which signal detection and impact analysis is carried out is important as this can influence the results. For instance, in the pilot study, impact analysis was performed using reactions that were detected at a particular preferred term. Occasionally, the higher-level term may have been more appropriate. Impact analysis could be performed at the higher-level term, but for this pilot study signals were detected and assessed at the preferred term level.

During the pilot phase we have occasionally encountered difficulties in estimating usage data, for example anticancer agents and over-the-counter products. On the occasions that it was not possible to estimate usage the default was used for the reporting rate input. Using Disability-Adjusted-Life-Years (DALYs) to estimate the non-fatal outcome is a possible further enhancement provided that these could be classified in the Medical Dictionary for Regulatory Activites (MedDRA) terminology.

We did not quantify the time taken to perform impact analysis, but it did not place an unacceptable burden on the staff involved. Its purpose is not necessarily to save time in the short term but to promote a more structured approach to prioritisation. It is important to recognise that our study does not allow any conclusions about the superiority of impact analysis or the collective judgement.

Conclusions

Impact analysis provides a useful approach to prioritising signals in the regulatory setting. It is a structured, repeatable method that provides a basis for the decision-making process and also encompasses a valuable audit trail. Its suitability for use by other regulatory agencies and the pharmaceutical industry would need to be investigated and possibly tailored to their needs.

Acknowledgements

We would like to thank the staff in the pharmacovigilance group at the Medicines and Healthcare products Regulatory Agency (MHRA) for their enthusiasm and patience throughout this project. We are grateful to Professor Stephen Evans for his discussions and Dr Lesley Wise and Mrs Eugenia Yoannou for their support and assistance. We also thank Peter Waller for his help in developing a computer program to facilitate impact analysis.

Funding was provided internally by the MHRA. The authors have no conflicts of interest that are directly relevant to the content of this study.

References

1. Waller P, Heeley E, Moseley J. Impact analysis of signals detected from spontaneous adverse drug reaction reporting data (Accompanying paper). Drug Saf 2005; 28 (10): 843-50
2. 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: 315-21
3. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999; 53: 177-89
4. Van Puijenbroek EP, Bate A, Leufkens HGM, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002; 11: 3-10
5. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001; 10: 483-6
6. Meyboom RH, Lindquist M, Egberts AC, et al. Signal selection and follow-up in pharmacovigilance. Drug Saf 2002; 25: 459-65
7. Bright RA, Nelson RC. Automated support for pharmacovigilance: a proposed system. Pharmacoepidemiol Drug Saf, 2002; 11: 121-5
8. van Puijenbroek EP, van Grootheest K, Diemont WL, et al. Determinants of signal selection in a spontaneous reporting system for adverse drug reactions. Br J Clin Pharmacol 2001; 52: 579-86
9. Waller PC, Lee EH. Responding to drug safety issues. Pharmacoepidemiol Drug Saf 1999; 8: 535-52
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74

Correspondence and offprints: Dr *Emma Heeley*, Room 15-153, Medicines and Healthcare products Regulatory Agency, Market Towers, 1 Nine Elms Lane, London, SW8 5NQ, UK.
E-mail: emma.heeley@mhra.gsi.gov.uk