Drug Safety 2008; 31 (9): 727-735
0114-5916/08/0009-0727/$48.00/0

A Model for Decision Support in Signal Triage

Bennett Levitan,[1] *Chuen L. Yee,*[2] *Leo Russo,*[3] *Richard Bayney,*[1] *Adrian P. Thomas*[3] and *Stephen L. Klincewicz*[3]

1 Pharmaceutical Portfolio & Decision Analysis, Johnson & Johnson Pharmaceutical Services, Titusville, New Jersey, USA

2 Medical Pharmacovigilance, Biotechnology/Immunology/Oncology Research & Development, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., Raritan, New Jersey, USA

3 Benefit Risk Management, Johnson & Johnson Pharmaceutical Research & Development L.L.C., Horsham, Pennsylvania, USA

Abstract

Spontaneous reporting of suspected adverse drug reactions (ADRs) has long been a cornerstone of pharmacovigilance. With the increasingly large volume of ADRs, regulatory agencies, scientific/academic organizations and marketing authorization holders have applied statistical tools to assist in signal detection by identifying disproportionate reporting relationships in spontaneous reporting databases. These tools have generated large numbers of signals defined as drug-ADR reporting associations that meet specified statistical criteria.

The challenge is to identify which signals are most likely to be medically important and therefore warrant priority for further investigation. Decisions related to signal triage are often complex and are based on a combination of clinical, epidemiological, pharmacological and regulatory criteria. There are no specific regulations, guidelines or standards that provide an objective basis for these decisions.

This paper describes preliminary work to identify and quantify the specific factors that contribute to a decision to prioritize a specific drug-ADR combination for further in-depth review. We applied a tool from the discipline of decision analysis to systematically assess the important attributes of spontaneously reported ADRs. A model was created that integrates these assessments and produces rankings for the signals generated from quantitative signalling methods. Although more research is necessary to evaluate the performance of this model fully, preliminary results suggest that the use of formal decision analysis approaches to support signal triage can provide potential benefit and will help meet an important need.

Spontaneous reporting of suspected adverse drug reactions (ADRs) has long been a cornerstone of pharmacovigilance. With the increasing overall volume of ADRs, the traditional practice of 'case by case' assessment by clinicians is increasingly suboptimal for signal detection. It has become necessary to develop quantitative methods/tools to assess data systematically at an aggregate level for detection of novel ADRs. Regulatory agencies, scientific/academic organizations and marketing authorization holders (MAHs) have applied statistical tools to assist in signal detection by identifying disproportionate reporting relationships in spontaneous reporting databases.[1-6] These tools have generated large numbers of signals defined as drug-ADR reporting associations that meet specified statistical criteria. However, upon further investigation, many of these signals have not turned out to be true ADRs. Of those that are confirmed, they are not equal in their clinical or public health importance.

The challenge in surveillance is to identify which signals are most likely to be medically important and therefore warrant priority for further investigation.[7,8] This triage function is often performed by a clinician based on a combination of clinical, epidemiological, pharmacological and regulatory criteria. It is a complex decision process that can lack transparency, is time-consuming and can be subject to various biases. There is little objective information and few methods to assess variability among or between assessors from MAHs, health authorities or scientific/academic organizations.

This paper describes preliminary work to identify and quantify the specific factors that contribute to a decision to prioritize a specific drug-ADR combination for further in-depth review, using a method known as multi-attribute decision analysis (MADA).

1. Methodology

1.1 Approach

When 'signals' are identified by quantitative signalling methods, the decision to prioritize a specific signal for further medical evaluation can be complex. This decision often involves an intuitive integration and weighting of many different types of data by an individual assessor. These data can be quantitative or qualitative. Quantitative data include reporting frequency and report volumes, categorical data (e.g. expectedness and biological plausibility), probabilities (e.g. disproportionality scores such as Empirical Bayes Geometric Mean [EBGM] and proportional reporting ratios) and time intervals (e.g. time to onset and time between events). Examples of qualitative data include clinical judgement of medical importance.

Unlike the causality assessment algorithms for individual safety reports and algorithms for signal detection from aggregate data, which have had extensive research and discussion,[2,3,6,7,9-11] there has been relatively little research on similar processes/tools for detecting and prioritizing signals from aggregate data.[12,13]

We conducted a review of the methods available for assessing and quantifying the decision-making process in other disciplines for potential applicability to the problem of signal triage and prioritization. Based on this review, we have identified a suitable approach from the decision-analysis literature called MADA, also known as multi-criteria decision analysis (MCDA).[14-16]

MADA modelling is a means to value and subsequently prioritize observations (ADRs or signals in this case) by taking into account the many characteristics by which they can be valued, often referred to as 'attributes'. When selecting a house, typical attributes might be cost, location and square footage. When considering different jobs, important attributes may be salary, job title and the length of the commute from home. This method involves building a hierarchy of attributes and graded scales for each attribute, rating each signal's performance on those attributes and then totalling the weighted scores for each signal. More complex approaches are also used, such as those used in this work. MADA models are particularly useful when there are multiple stakeholders, each of whom emphasizes different issues when ranking an observation. These models can handle both qualitative and quantitative data and, when properly designed, can account for missing information and ambiguous entries. Outputs of the model enable rankings according to the relative performance of all observations on various subsets of the input attributes, and allow for explanations as to why one signal is ranked above another. For example, a MADA model would make it clear that ADR *X* is ranked higher than ADR *Y* because of

Table I. Attributes and their definitions

Attribute	Definition
% Serious cases	Absolute percentage (0–100%) of cases meeting the US FDA definition of 'serious' for a given AE in the Johnson & Johnson safety database
Confounded by indication[a]	Is the AE potentially attributable to indications for which the drug is typically used?
Drug class effects[a]	Has this AE been reported in other drugs in its class?
Empirical Bayes Geometric Mean[b]	A measure of disproportional reporting for a given AE
Expectedness[a]	Is this AE listed in the company core data sheet or equivalent?
External interest[a]	Is there interest from any of the following categories: media, health authorities, medical/ scientific community or legal?
Fractional reporting ratio	Ratio of reporting fraction from two different time periods; a measure of interval change in reporting for a given AE
Positive rechallenge	Is there at least one report of a positive rechallenge associated with this signal within the Johnson & Johnson safety database?
Targeted medical events	Matching to a constructed list of targeted medical events composed of events described in the FDA proposed rule on *Safety Reporting Requirements for Human Drug and Biological Products* (14 March 2003),[18] SAEs in 'Dear Doctor' letters and events identified as preventable (medication error, accidental overdose, drug interaction)
Typical of ADRs	Does the given AE match to a constructed list of AEs, which are typical of being ADRs?
Volume of reports	The cumulative number of reports of the AE up to and including the current reporting period within the Johnson & Johnson safety database

a Determined by SSPs based on medical judgement. All other attributes were computed from the adverse event reporting system or Johnson & Johnson databases.

b WebVDME, a commercial data-mining tool by Phase Forward.

ADRs = adverse drug reactions; **AE** = adverse event; **SAEs** = serious adverse events; **SSP** = safety surveillance physician.

a much larger societal impact, even though it is less likely to be causally related to the drug. Additional advantages of the MADA approach are that the models are easily understood, the process for developing and applying them is transparent and they are easy to modify to novel situations, such as when considering a different class of drugs or patient population.

1.2 Developing the Model

The objective of the model was to generate systematic rankings (numeric scores) to support signal triage decision-making in Johnson & Johnson Benefit Risk Management's surveillance process. During triage, safety surveillance physicians (SSPs), the physicians who are responsible for routinely reviewing surveillance data,[17] review a list of signals (ADRs meeting a set of defined criteria, e.g. statistical threshold) and make individual triage decisions to prioritize them for further investigations.

Building the MADA models involved structured interviews with the different decision-makers and stakeholders. These stakeholders included other physicians and scientists involved in assessing and managing postmarketing safety issues (referred to as 'benefit-risk leaders' and their supervisors, 'therapeutic area leaders') and the epidemiologists who support postmarketing safety assessment.

The interviews were conducted in order to:

Table II. Three key objectives under which all adverse drug reaction (ADR) attributes were grouped

Key objective	Description
Novelty of event	Degree to which a signal has not been observed before or the degree to which the reporting frequency of a known ADR has changed
Strength of evidence	Degree to which we believe the ADR represents a causal relationship between the drug and the event
Medical impact	Degree to which such a causal relationship impacts on patients' lives or how the potential relationship is viewed by regulatory bodies/scientific community

1. identify the concepts used when making prioritization decisions (objectives);
2. show the hierarchical relationships between these objectives (objective hierarchy);
3. identify properties of the ADRs to assess or approximate these objectives (attributes);
4. elicit relationships between the attributes and how they may influence decisions (utility functions);
5. elicit weights that reflect the relative importance of the different attributes.

In initial interviews, a candidate set of >30 objectives was refined into a set of 11 key attributes used by SSPs during triage. The selected attributes were based on data generally available in spontaneous ADR reports and had to be measurable at the time the model was applied. These are summarized in table I.

Additional discussion led to grouping of these 11 attributes under three key objectives: novelty of event, strength of evidence and medical impact (table II and figure 1). Conceptually, these key objectives are similar to the 'SNIP' criteria (i.e. the strength of the signal, whether it is new, clinically important or whether there is potential for preventative measures) described by Waller and Lee,[8] and to components of the impact analysis approach by Waller et al.[12] and Heeley et al.[13]

In most cases, the attributes were well accepted measures of the objectives; for example, a positive rechallenge is strong evidence supporting a drug association. In other cases, proxy attributes were necessary, such as using the percentage of spontaneous reports categorized as 'serious' as a proxy for the medical impact of an adverse event to the patient.

For continuous attributes such as EBGM, we developed utility functions through structured team interviews using the 'midrange' technique.[14,15] Binary attributes received utilities of 0 and 1 for the two values. These functions converted each attribute's measured value into a normalized measure of 'priority for investigation' and allowed us to capture the decision-maker's sense of where the important changes in the attribute occur. Finally, each attribute and objective was weighted to compute overall rank. Weights were elicited in structured interviews but were subsequently revised, as described later.

During preliminary use of the model, informal comparison between rankings generated by the model and those by the SSPs showed that the

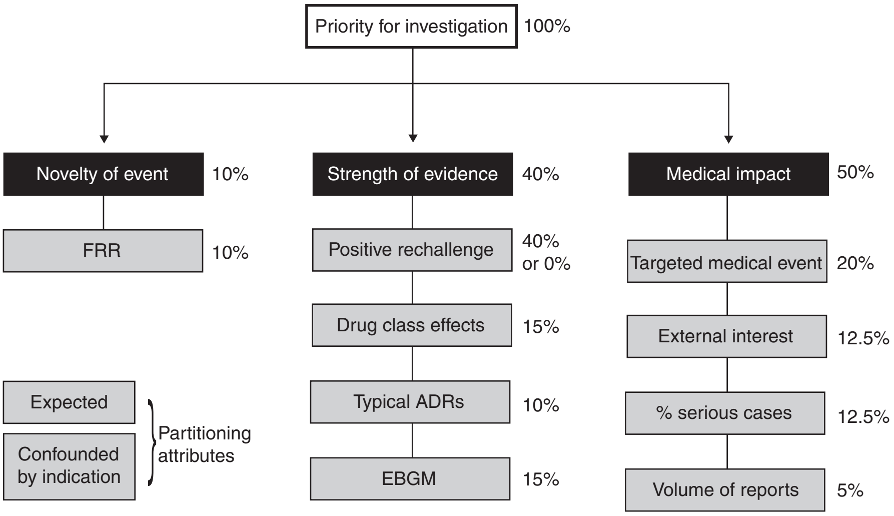

Fig. 1. Adverse drug reaction (ADR) prioritization model: objective hierarchy. Attributes were grouped into three key objectives (shown by the black boxes and described in table II). Weights shown are for the revision 1 model. The attributes 'expected' and 'confounded' were originally included under 'novelty of event' and 'strength of evidence', respectively, but are now used as shown in figure 2. **EBGM** = Empirical Bayes Geometric Mean; **FRR** = fractional reporting ratio.

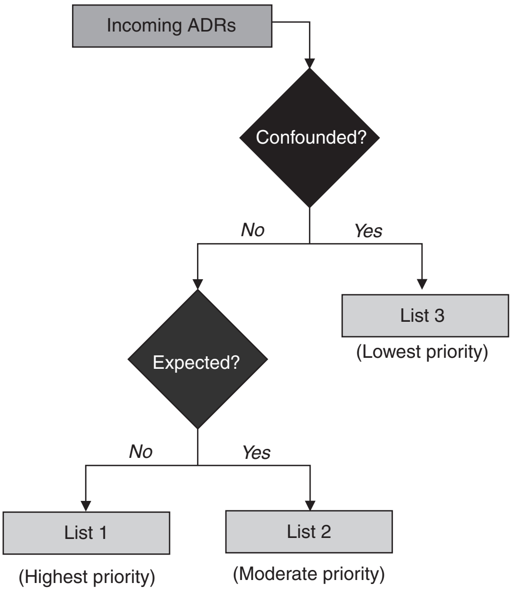

Fig. 2. Adverse drug reaction (ADR) prioritization model: partitioning logic for ADRs. The incoming ADRs were partitioned into three subsets: list 1 = unconfounded, unexpected ADRs (highest priority); list 2 = unconfounded, expected ADRs (moderate priority) and list 3 = confounded ADRs (lowest priority).

clinicians were using two attributes of ADRs differently from what had been expected initially. Investigation of ADRs confounded by indication and those already listed in the compound's company core data sheet was consistently judged to be of lower priority than for those without these properties, beyond what could fit into the MADA framework. To address this finding, the ADRs were partitioned into three subsets (lists 1, 2 and 3; figure 2), then ranked within each set by multi-attribute models. Since the unconfounded and unexpected ADRs in list 1 often have the highest priority for investigation, our analysis focuses on this subset of ADRs identified during the signalling process.

This partitioning also entailed revisiting the weights initially elicited for the full set of ADRs. To accommodate the subset of ADRs that comprise list 1, the model weights were revised and are referred to as 'revision 1'. Most of the analyses are based on revision 1 weights. We also considered a third set of weights, 'revision 2', which incorporated the idea that list 1 ADRs are implicitly novel and need little weight on novelty of event and thus more heavily favoured medical impact attributes.

In section 3, we comment on the importance of including list 2 and 3 ADRs in any application of this approach as well as on the implications of revising the weights after preliminary use.

1.3 Testing the Model

To be representative of the company product portfolio, we selected three compounds that differed in their stage of product life cycle and their therapeutic class (one compound each for the treatment of cancer, neurological disorders and disorders of the immune system/connective tissue and joints). A test dataset consisting of approximately 75 ADRs and the associated attribute scores for each ADR was created for each compound. For each drug, the ADR list was generated from the union of three sources: the top 25 ADRs by EBGM score using the US FDA Adverse Event Reporting System (AERS) database; the top 25 by the fractional reporting ratio; and the top 25 by report volume from the company safety database. The size of the final ADR list varied as a result of overlapping between these sources. Attribute information was obtained from a combination of three sources: FDA AERS, the company safety database and clinical judgement by the SSPs. The SSPs were provided with guidelines for those attributes based on clinical judgement (confounding by indication, drug class effect, expectedness and external interest). A tester group of physicians was instructed to rank the ADR list in order of priority for investigation and to provide their supporting rationale. We compared the prioritization rankings from the physicians with those generated by the model.

2. Results and Analysis

We first compared SSP tester prioritizations with those from the model. The tester group consisted of three SSPs, all of whom were involved in developing the model. The two key measures used were the Spearman rank correlation coefficient[19] and correlation plots. Spearman rank correlation coefficient is a measure of the degree to which two rankings

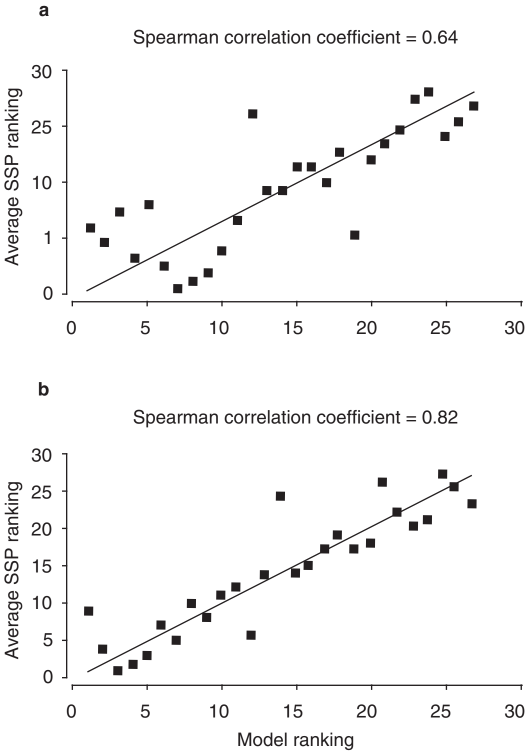

Fig. 3. Correlation plots for drug 1 using: (**a**) original weights; (**b**) revision 2 weights. The lines show where the markers would lie with perfect correlation. **SSP** = safety surveillance physician.

differ. The larger the number of items ranked differently and the greater the degree to which each item's position differs in the two rankings, the lower the correlation. A correlation coefficient of 1 means an exact match in ranking, a correlation coefficient of 0 means no correlation and a correlation coefficient of −1 indicates complete anti-correlation, in which the two rankings are in opposite order. Since this measure can be abstract, we also show comparisons between rankings by using correlation plots. Correlation plots use markers for each item being ranked, with the x-axis being model rankings and the y-axis being the average SSP rankings. Perfect correlation (Spearman rank correlation coefficient = 1.0) results in all items lying on the diagonal.

For drug 1, there were 27 unique ADRs in list 1, for drug 2 there were 8, and for drug 3 there were 12. For a combined ranking, each ADR was assigned the average rank amongst the three testers. For the three drugs, the correlation plots suggest modest to good correlation. Figure 3a shows the correlation plot for drug 1 with the original weights. The correlation plots for drugs 2 and 3 are available as supplementary material online ('ArticlePlus') at http://drugsafety.adisonline.com. The effect of using this model was compared with models based on the original weights elicited for the model and with increased relative weight for medical impact attributes (figure 3b and table III).

We then expanded the tester group to 12 physicians from two safety functional groups: six SSPs (who have responsibility for routine review of surveillance data) and six benefit-risk leaders (who have overall responsibility for drug safety and benefit-risk for a compound) using the same drugs. Most testers were not involved in developing the model. In addition, testers were allowed to use equivalence ranking, indicating that two or more ADRs had the same priority for investigation.

To minimize potential masking of inter-physician variability by using average SSP rankings, we compared median tester rankings with model rankings using both Spearman rank correlation coefficient and Kendall's tau (τ) statistic. Similar to Spearman rank correlation coefficient, Kendall's τ quantifies the level of agreement in ordering between two sets of rating.[20] As is typically the case, both statistics showed similar results[21] with consistently higher values for Spearman rank correla-

Table III. Correlation between model and safety surveillance physicians: effect of relative weight of attributes for list 1

Version	Attribute weight (%)			Spearman correlation coefficient		
	novelty	strength of evidence	medical impact	drug 1	drug 2	drug 3
Original	10	50	40	0.64	0.69	0.51
Revision 1	10	40	50	0.78	0.69	0.51
Revision 2	5	25	70	0.82	0.69	0.65

Table IV. Correlation between model and testers using Spearman rank correlation coefficient and Kendall's τ statistic. Spearman and Kendall results show the same patterns for all drugs and groups of testers

Drug	Spearman	Kendall
Drug 1	0.85	0.67
Drug 2	0.86	0.79
Drug 3	0.52	0.39
Overall	0.77	0.62

tion coefficient (table IV). Correlation was quite good for the first two drugs and less so for the third.

Our final set of results concern inter-tester agreement as a means to gauge how close a correlation with testers is theoretically possible for the model. Figure 4 shows Kendall's coefficient of concordance (*w*) for the SSP and benefit-risk leader groups. Kendall's coefficient quantifies the degree of agreement between multiple raters and ranges from 0 for no agreement to 1 for perfect agreement.[20] According to Landis and Koch's[22] criteria for agreement, the results show mostly moderate agreement between testers, with no clear difference between the two physician groups tested. The scores also suggest that the degree of variability between testers increases as the number of ADRs increases.

Because the prioritization model was not designed to match any given tester, we expected tester-tester correlation to be greater than model-tester correlation. It was surprising to see that the correlations were comparable. Using the tester/tester and tester/model correlation results alone, an outsider to this experiment would have difficulty distinguishing the model's ranking from that of a human tester.

3. Discussion

Decisions related to signal triage are often complex and there are no specific regulations, guidelines or standards that provide an objective basis for these decisions. We applied a well described tool from the discipline of decision analysis to assess systematically the important attributes of spontaneously reported ADRs. A model was created that integrates these assessments and produces rankings for the generated signals from quantitative signalling methods.

Our research to date focused on developing and testing a prioritization model that dealt with unconfounded, unexpected ADRs (list 1). In any real-world application of this approach to signal triage, it will be critical to also have prioritization models for list 2 (unconfounded, expected) and list 3 (confounded) ADRs. While ADRs in list 1 are more likely to be medically important signals warranting high priority for further investigation, important signals may also come from list 2 or 3. For example, relying solely on list 1 would risk missing important signals that are confounded by indication, e.g. drug-induced hepatitis in a drug used for the treatment of hepatitis. There will also need to be a means to merge the prioritization rankings from the three lists or to develop algorithms or 'trigger thresholds' for when ADRs in lists 2 and 3 require further investigation. These are potential topics for future research.

A key observation from the preliminary results is that correlation between the model and the testers as a group (either as average or median tester ranking) was consistently higher than correlation between the testers or between the model and the individual tester. This suggests that the model is more representative of the thinking of the group rather than that

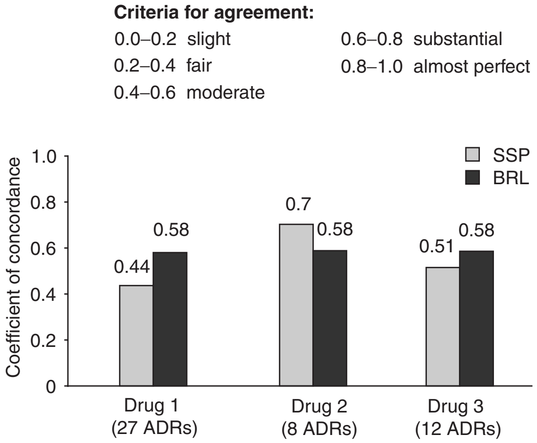

Fig. 4. Degree of inter-tester variability according to Kendall's coefficient of concordance. 'Criteria for agreement' are from Landis and Koch[22] and provide some subjective guidance in interpreting the scales. The scores suggest that the degree of variability between testers increases as the number of adverse drug reactions (ADRs) increases. **BRL** = benefit-risk leader; **SSP** = safety surveillance physician.

of the individual physicians. Since the model was developed by a group and is intended to help physicians perform signal triage in a consistent and transparent manner, this is a very encouraging result. However, inter-physician variability may be a useful measure in signal triage. This should to be explored further and a measure of inter-physician disagreement could be developed and added in future model refinement.

Another observation is that for unconfounded and unexpected ADRs, physicians appear to place more weight on medical impact attributes than strength of evidence types of attributes in their prioritization. However, since only three drugs and a small number of assessors were used in the testing thus far, this will need to be evaluated further with a larger number of drugs and, ideally, a larger and more heterogeneous group of assessors. Further revisions of attribute weights may be necessary to optimize model performance.

Two important caveats to the work are consequences of the limited number of drugs used in the test dataset during model development and testing. In developing the MADA, preliminary results diverged enough from tester results that it became clear we needed to partition the ADR list into three subsets of ADRs (list 1–3). A different model is needed for each list, differing in attribute weights since clinical judgement will affect triage differently for the different types of ADRs in each list. In addition, the weights of the attributes were revised to favour medical impact based on preliminary results. Since the same three drugs were used in 'tuning' and testing the model, the results reported are not independent of the model-design process. For the same reasons, until we get a larger test-drug dataset, it is not possible to demonstrate the degree to which the method would perform across compounds. We plan to address these limitations in future work with additional compounds across therapeutic classes.

In the next phase of this project, we will evaluate the potential value of implementing the signalling triage model in the current surveillance process. The SSPs will review scheduled surveillance reports in accordance with current process and document the ADRs identified for further review. They will then be provided with output from the model including the ADR attribute scores for the highest ranked ADRs identified by the model for further review. The SSPs will be instructed to compare their list of ADRs for further review with those identified by the model and to comment on any differences. In addition, they will be asked to document whether the review of the model rankings resulted in a revision of their original list of ADRs for further review. Any change in SSP decision is potential value added by the model. The collected data will be reviewed to determine the need for model refinement to optimize concordance with the SSPs. It is important to reiterate that the model is intended to provide support to the physician in making signal triage decisions and is not intended to replace the physician or clinical judgement.

During the next phase, we will target and collect data involving approximately 10–12 medicinal products. With a larger and more diverse group of products, the model can be tested for generalizability across compounds. As noted in the description of the 11 attributes above, some of these are not straightforward and required determination by the SSPs based on medical judgement (confounding by indication, drug class effect, expectedness according to the company core data sheet, and external interest). In order to expand the number of medicinal products for testing, automating the data generation for some of these attributes will be required. We are still exploring different approaches to determine how best to address this.

We plan to adopt an iterative approach where we will evaluate the model on a regular basis, e.g. annually, and determine the need for model refinement based on the collected data. This approach would be similar to obtaining a posterior probability in a Bayesian approach.[23] This would also allow us to assess performance of the model in a real-world environment.

While more research is necessary to evaluate the performance of this model fully, preliminary results suggest that the use of formal decision analysis approaches to support signal triage can provide potential benefit and will help meet an important need.

Acknowledgements

All authors were full-time employees of Johnson & Johnson at the time of the completed work. No sources of funding were used in the preparation of this article. The authors have no conflicts of interest that are directly relevant to the content of this article. We would like to acknowledge the important contributions of other project team members to this work: Karen Kaplan, Rezaul Karim, Kenneth Kwong, Anders Lindholm, Aparna Mohan, Stanley Music, Mary Panaccio and Saurabh Raman.

References

1. Bate A, Linquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998; 54: 315-21
2. 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
3. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 2002; 25 (6): 381-92
4. Eudravigilance Expert Working Group. Draft Guideline on the use of statistical signal detection methods in the Eudravigilance data analysis system. London: European Medicines Agency; 2006 Nov 16. EMEA/106464/2006
5. Purcell P, Barty S. Statistical techniques for signal generation: the Australian experience. Drug Saf 2002; 25 (6): 415-21
6. Almenoff J, Tonning A, Gould L, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf 2005; 28 (11): 981-1007
7. Stahl M, Lindquist M, Edwards IR, et al. Introducing triage logic as a new strategy for the detection of signals in the WHO drug monitoring database. Pharmacoepidemiol Drug Saf 2004; 13: 355-63
8. Waller P, Lee E. Responding to drug safety issues. Pharmacoepidemiol Drug Saf 1999; 8 (7): 535-52
9. Clark JA. Algorithms. In: Mann RD, Andrews EB, editors. Pharmacovigilance. 1st ed. Chichester: John Wiley & Sons, 2002: 229-46
10. Holden WL, Juhaeri J, Dai W. Benefit-risk analysis: a proposal using quantitative methods. Pharmacoepidemiol Drug Saf 2003; 12: 611-6
11. Holden WL, Juhaeri J, Dai W. Benefit-risk analysis: examples using quantitative methods. Pharmacoepidemiol Drug Saf 2003; 12: 693-7
12. Waller P, Heeley E, Moseley J. Impact analysis of signals detected from spontaneous adverse drug reaction reporting data. Drug Saf 2005; 28 (1): 843-50
13. Heeley E, Waller P, Moseley J. Testing and implementing signal impact analysis in a regulatory setting: results of a pilot study. Drug Saf 2005; 28 (10): 901-6
14. Kirkwood CE. Strategic decision making: multiobjective decision analysis with spreadsheets. Belmont (CA): Duxbury Press, 1996
15. Keeney RL. Value-focused thinking: a path to creative decision-making. Cambridge (MA): Harvard University Press, 1996
16. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. Cambridge: Cambridge University Press, 1993
17. Yee CL, Klincewicz SL, Knight JF, et al. Practical considerations in developing an automated signaling program within a pharmacovigilance department. Drug Inf J 2004; 38: 293-300
18. Safety reporting requirements for human drug and biological products: proposed rule. Federal Register 2003 Mar 14; 67 (50) [online]. Available from URL: http://www.fda.gov/Cber/rules/safereport.htm [Accessed 2008 Jul 8]
19. Hogg R, Craig A. Introduction to mathematical statistics. 5th ed. Englewood Cliffs (NJ): Prentice Hall, 1994
20. Kendall MG. Rank correlation methods. 4th ed. London: Griffin, 1970
21. Siegel S, Castellan N. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill, 1988
22. Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977; (33): 159-74
23. Landrum MB, Normand ST. Applying Bayesian ideas to the development of medical guidelines. Stat Med 1999; 18: 117-37

Correspondence: Dr *Chuen L. Yee*, Medical Pharmacovigilance, Biotechnology/Immunology/Oncology Research & Development, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 920 Route 202 South, Raritan, NJ 08869, USA.
E-mail: cyee@prdus.jnj.com