ORIGINAL RESEARCH ARTICLE

Drug Saf 2010; 33 (11): 1045-1054
0114-5916/10/0011-1045/$49.95/0

Comparison of Three Methods (Consensual Expert Judgement, Algorithmic and Probabilistic Approaches) of Causality Assessment of Adverse Drug Reactions

An Assessment Using Reports Made to a French Pharmacovigilance Centre

Hélène Théophile, Yannick Arimone, Ghada Miremont-Salamé, Nicholas Moore, Annie Fourrier-Réglat, Françoise Haramburu and *Bernard Bégaud*

Université de Bordeaux, Département de Pharmacologie, INSERM U657, Bordeaux, France

Abstract

Background: Different methods have been proposed for assessing a possible causal link between a drug treatment and an adverse event in individual patients. They approximately belong to three main categories: expert judgement, operational algorithms and probabilistic approaches.

Objective: To compare, in a set of actual drug adverse event reports, three different methods for assessing drug causality, each belonging to one of the three main categories: expert judgement, the algorithm used by the French pharmacovigilance centres since 1985, and a novel method based on the logistic function.

Methods: Fifty drug-event pairs were randomly sampled from the database of the Bordeaux pharmacovigilance centre, France. To serve as the gold standard, the probability for drug causation, from 0 to 1, was first determined for each drug-event pair by a panel of senior experts until consensus was reached. Causality was then assessed by members of the Bordeaux pharmacovigilance centre by using the French algorithm and the logistic method. Results expressed as a probability with the logistic method and as a score from 0 to 4 with the French algorithm were then compared with consensual expert judgement, as were the sensitivity, specificity and positive and negative predictive values.

Results: Probabilities ranged from 0.08 to 0.99 (median 0.58; mean 0.60) for experts versus 0.18–0.88 (median 0.73; mean 0.67) for the logistic method. Consensual expert judgement was not discriminant ($p = 0.50$) in ten cases. For the algorithm, only three of five causality scores were found, *doubtful* scores being clearly predominant (74%) followed by *possible* (16%) and *probable* (10%) scores. Sensitivity and specificity were 0.96 and 0.42, respectively, for the logistic method versus 0.42 and 0.92 for the algorithm. Positive and

negative predictive values were 0.78 and 0.83, respectively, for the logistic method versus 0.92 and 0.42 for the algorithm.

Conclusions: Agreement between the three approaches was poor, and only satisfactory for drug events judged as drug-induced by consensual expert judgement. The logistic method showed high sensitivity at the expense of poor specificity. Conversely, the algorithm had poor sensitivity but good specificity. The comparatively good sensitivity and positive predictive values of the logistic method suggest that it may be more useful in the routine or automated assessment of case reports of suspected but still unknown adverse drug reactions. With a substantial rate of false positives relative to true negatives (low specificity), the logistic method does not replace, but can be complemented by, critical clinical assessment of individual cases in evaluating drug-related risk.

Background

More than 20 different methods have so far been proposed for assessing a possible causal link between a drug treatment and an adverse event in individual patients. They approximately belong to three main categories: expert judgement, probabilistic approaches and operational algorithms.[1-3]

Expert judgement or global introspection, which takes into account all available and relevant data on the case considered, sets out to mimic the clinical diagnostic process. Consequently, it generally suffers from both subjectivity and lack of standardization, leading to a poor reproducibility and to intra- and inter-rater disagreements.[4-7] These limitations can be circumvented if assessment is made by several senior experts interacting on a consensus basis, e.g. by using the Delphi method.[8] Although relatively complex, this approach is often considered the gold standard for causality assessment.

Most published probabilistic approaches have been derived from Bayes' theorem. They have several indisputable advantages,[9] in particular (i) the basic rule of probability theory is respected, since the absence of any relevant information, either by lack of information or by the absence of conclusive arguments for or against drug responsibility, should lead to a neutral estimate, i.e. a probability of 0.5 or an odds of 1;[10] and (ii) a precise estimate of drug causation, formalized as probability or odds, is obtained on a continuous scale. However, formal Bayesian approaches are troublesome to use in routine practice because they require precisely quantified information to model probability distributions for each parameter, even if, in certain cases, assumptions can be made.[10,11]

The algorithmic approach consists of successive assessment of criteria under binary or multiple-choice form, combined with a sum of scores or decision tree. This approach is convenient, easy to use and said to minimize inter- or intra-observer variability. However, final assessment produced by a given algorithm depends highly on the relative weight of each criterion, which is fixed more or less arbitrarily by the author(s) of the method.[10,12] As a result, in routine practice when consensual experts' judgement is out of reach, algorithms are by far the most commonly used, even if Bayesian approaches are much more satisfactory, albeit difficult to use.

A novel approach based on the logistic function[13] has recently been proposed to model consensual expert judgement. The method conserves the basic principles and simplicity of algorithms, but expresses the causality assessment as a probability, which can be directly obtained from a computerized version of the method.

The aim of the present study was to compare this novel method, an operational algorithm, and consensual expert judgement taken as reference. The algorithm used was the one routinely used by

the French pharmacovigilance network to assess more than 200 000 cases of adverse drug reactions (ADRs).[14]

Methods

Fifty adverse events randomly sampled among those reported to the Bordeaux pharmacovigilance centre were considered. For each case, available information in the complete file was summarized in a standardized form, including the characteristics of the patient, the suspected drug(s) with the dates of treatment, the type of event, the date of event onset, relevant biological and clinical data, other current medical treatments and the time course of the event. This study sample included 39 different types of ADRs distributed in 13 organ classes, the most often reported being CNS (14/50 [28%]) and skin (9/50 [18%]), followed by blood (7/50 [14%]), body as a whole (6/50 [12%]) and liver/pancreas disorders (5/50 [10%]). The proportion of cases qualified as serious according to the WHO criteria was 50% in the study sample (19 hospitalizations, 4 life-threatening reactions, 1 death and 1 case of sequelae). Of the 50 cases, 31 were referred from university hospitals, 9 from other hospitals or clinics and 10 from private healthcare professionals.

The 50 drug-event pairs were assessed separately by members of the Bordeaux pharmacovigilance centre and by experts (see below).

Evaluation using Consensual Expert Judgement

Consensual expert judgement was used as the gold standard for drug causation. A panel of 26 experts (7 specialized in clinical pharmacology and 19 in internal medicine) was set up to evaluate the causality assessment of the 50 drug-event pairs by using expert judgement. For each pair, a group of three experts (one senior in pharmacovigilance and two clinicians) was selected. The clinicians were chosen according to their field of expertise either with regard to the type of event, e.g. hepatologist for hepatitis, or the treated disease, e.g. a haematologist reviewed a case of seizures in a patient treated with vincristine for acute lymphoblastic leukaemia. Each of the three experts was asked to express separately his/her judgement on the likelihood that the suspected drug was causal based on a 100 mm visual analogue scale (VAS). The judgement was then directly converted into a probability of drug causation ranging from 0 to 100% or 0 to 1.0. Three possibilities were considered: (i) when the three experts were in total agreement, i.e. judgements differed by ≤5 mm on the VAS, the probability was accepted; (ii) when the extreme judgements differed by >5 mm but ≤25 mm, the arithmetic mean of the three scores was used; and (iii) when the difference between the extreme judgements was >25 mm, the drug-event pair and expert judgement were reanalysed on a consensus basis by a second group of four senior experts (two seniors in pharmacovigilance, two clinicians), independent of the first group; the final probability was accepted.

To comply with current practice of drug causality assessment, the likelihood of a causal relationship for each of the 50 drug-event pairs was assessed by the senior staff (two physicians and three pharmacists) from the Bordeaux pharmacovigilance centre, using the French algorithm and the logistic method.

Evaluation of Drug Causation using the Algorithm

The French algorithm,[15] detailed in table I, is based on evaluation of seven criteria belonging to two groups: chronological and semiological. The assessment of the chronological criteria results in a score ranging from 0 to 3: C0 (*incompatible*), C1 (*dubious*), C2 (*possible*) or C3 (*likely*). The evaluation of semiological criteria gives a score ranging from 1 to 3: S1 (*dubious*), S2 (*possible*) or S3 (*likely*). The combination of the chronological and semiological scores results in the so-called *intrinsic score* from 0 to 4 corresponding to 5 degrees: *not related*, *doubtful*, *possible*, *probable* and *highly likely*, respectively (table II). The term *intrinsic* means that imputation only considers information from the case without direct reference to the fact that it was previously reported or published, even if the 'extrinsic' score (a score indicating how well known the adverse reaction

Table I. Distribution of criteria of causality assessment and associated scores in the French algorithm[15]

Criteria investigated	Scores
Chronological criteria (C)	
Time to onset	
Incompatible	
Compatible	
Highly suggestive	
Dechallenge	
Against the role of the drug	
Not conclusive or not available	
Suggestive	C0–C3
Rechallenge	
Negative	
Not available or not conclusive	
Positive	
Semiological criteria (S)	
Search for non-drug-related causes	
Not investigated and/or possible non-drug cause	
Absent	
Evocative semiology of drug responsibility and/or risk factor(s) for drug reaction	S1–S3
Absent or ruled out	
Present and well validated	
Specific validated laboratory test	
Negative	
Not made or not available	
Positive	
Bibliographical score (B)	
Reaction never reported or not published	
Not well known, previously published only once or twice	B0–B3
Well known reaction	
Reaction not fulfilling above definitions	

is, ranging from 0, 'never reported' or 'not published', to 3, 'well known ADR') can be added.

Evaluation of Drug Causation using the Logistic Method

The logistic method, as detailed in table III, combines seven causality criteria.[13] Its main purpose is to reproduce consensual expert judgement as much as possible. To do so, the weight associated with each criterion is computed by using multilinear regression, also with a consensus of experts as reference. The final causality assessment for a given case is determined by the sum of scores of the seven criteria, which is then transformed by logistic function into the probability of drug causation ranging from 0 to 1 (figure 1). As for Bayesian approaches, absence of relevant or discriminant information leads to a probability of 0.5, probabilities higher than 0.5 being in favour of drug responsibility, and those lower than 0.5 not in favour of it.

Statistical Analysis

Descriptive analysis was based on median, mean, range and quartiles for probabilities, i.e. for consensual expert judgement and logistic method, and on distribution of causality scores for the French algorithm. The nonparametric Mann-Whitney U test was used to compare the probabilities of drug causation between experts. To calculate the correlation between the different methods, the continuous variables expressed as probabilities were converted into ordinal variables as follows: probabilities from 0 to 0.05, *not related*; 0.05–0.40, *doubtful*; 0.40–0.60, *possible*; 0.60–0.95, *probable*; and 0.95–1, *highly likely*. The correlation between the logistic method and consensual expert judgement on one hand, and the French algorithm and consensual expert judgement on the other hand, was assessed by the Spearman's correlation coefficient. Sensitivity, specificity and positive and negative predictive values of both the logistic method and the French algorithm were computed with reference to the gold standard obtained from consensual expert judgement. This part of the analysis was restricted to the 38 cases for which consensual expert judgement was clearcut, i.e. the 12 with a probability of the drug being responsible of <0.45 and

Table II. Decision table for the intrinsic imputability in the French algorithm.[15] Shown are scores for chronological and semiological imputabilities[a]

Chronology	Semiology		
	S1	S2	S3
C0	0	0	0
C1	1	1	2
C2	1	2	3
C3	3	3	4

a The drug-effect relation can be 0 *not related*, 1 *doubtful*, 2 *possible*, 3 *probable* and 4 *highly likely*. See table I for further details.

Table III. Distribution of criteria of causality assessment and associated scores in the logistic method[13]

Criteria	Scores
Time to onset	
Incompatible	
Not suggestive	X_1
Unknown or not available	
Compatible	
Highly suggestive	
Dechallenge	
Against the role of the drug	
Not conclusive or not available	X_2
Suggestive	
Rechallenge	
Negative	
Not available or not conclusive	X_3
Positive	
Search for non-drug-related causes	
Non-drug cause highly probable	
Not investigated and/or possible non-drug cause	X_4
Non-drug cause ruled out	
Risk factor(s) for drug reaction	
Ruled out or absent	X_5
Well validated and present	
Reaction at site of application or plasma concentration known as toxic, or validated laboratory test	
Unrelated or not available	X_6
Present and/or positive	
Previous information on the drug and symptomatology	
Reaction not previously reported	
Not available	X_7
Not well known, previously published only once or twice	
Labelled reaction	
Sum of scores	ΣX_i

the 26 with a probability of >0.55. Results of the whole sample were also presented in order to test the influence of the 12 non-discriminant assessments on the behaviour of the logistic method and the French algorithm. For the logistic method, a probability of ≤0.50 was considered as not being in favour of the responsibility of the drug and a probability of >0.50 as being in favour of it. For the French algorithm, the scores 0 (*not related*) and 1 (*doubtful*) were considered as negative and scores 2 (*possible*), 3 (*probable*) and 4 (*highly likely*) as positive.

Computations were made using the SAS package (version 8.1; SAS Institute, Cary, NC, USA).

Results

An agreement between the three experts using separate global introspection was obtained in only 13 cases. Therefore, a new consensual assessment was necessary for 74% (37/50) of the cases. The probabilities of drug causation for the 13 cases for which a consensus was obtained straightaway were higher than those of 37 cases for which a second evaluation was necessary, although not to a statistically significant extent (median 0.78 vs 0.50; $p = 0.12$).

Causality assessments given by the three approaches for the 50 drug-event pairs are presented in table IV. The median probability for drug causation for the 50 events was 0.58 for experts (table V) with a range from 0.08 to 0.99, a first quartile at 0.48 and a third quartile at 0.84. The mean was 0.60 (SD 0.25). For the logistic method (table V), the median probability was 0.73 with a range from 0.18 to 0.88, a first quartile at 0.60 and a third quartile at 0.73. The mean was 0.67 (SD 0.15). For the algorithm, the distribution of causality scores was clearly skewed to the left, with the score 1 (*doubtful*) being clearly predominant as it occurred in 37 of the 50 cases (74%) as a result of combinations of low chronological and semiological scores. For 19 of these 37 cases (51%), the temporal relationship between drug treatment and the event was considered as poor (C1) and for 34 (92%), no suggestive clinical or biological pattern was present, or an alternative cause was suspected or present (S1). Moreover, only three of the five causality levels were obtained, the two extremes,

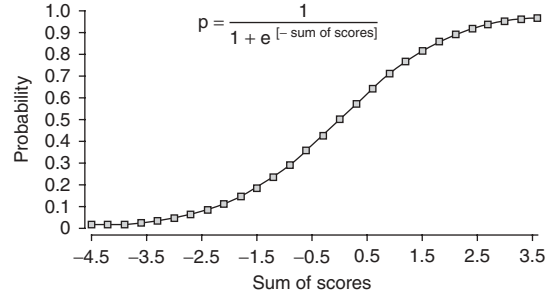

Fig. 1. Obtaining the probability of drug causation from the sum of scores by using the logistic function.

Table IV. Causality assessment given by the three approaches for the 50 drug-event pairs

Case no.	Experts' probability	Logistic probability	French algorithm intrinsic score[a]	Case no.	Experts' probability	Logistic probability	French algorithm intrinsic score[a]
1	0.08	0.18	1 (C1S1)	26	0.60	0.60	1 (C1S1)
2	0.10	0.48	1 (C1S1)	27	0.64	0.73	1 (C2S1)
3	0.20	0.50	1 (C1S1)	28	0.65	0.73	1 (C2S1)
4	0.20	0.71	1 (C2S1)	29	0.65	0.82	1 (C2S1)
5	0.22	0.45	1 (C1S1)	30	0.65	0.73	1 (C2S1)
6	0.25	0.67	1 (C1S1)	31	0.75	0.73	1 (C2S1)
7	0.25	0.45	1 (C1S1)	32	0.75	0.62	1 (C2S1)
8	0.27	0.80	1 (C2S1)	33	0.78	0.62	1 (C1S2)
9	0.30	0.60	1 (C1S1)	34	0.80	0.62	1 (C1S1)
10	0.35	0.71	1 (C2S1)	35	0.80	0.73	2 (C2S2)
11	0.40	0.87	1 (C2S1)	36	0.80	0.87	2 (C2S2)
12	0.40	0.87	2 (C2S2)	37	0.80	0.82	3 (C3S2)
13	0.48	0.57	1 (C2S1)	38	0.85	0.79	3 (C3S1)
14	0.48	0.52	1 (C1S1)	39	0.85	0.87	2 (C2S2)
15	0.50	0.29	1 (C1S1)	40	0.86	0.73	1 (C2S1)
16	0.50	0.60	1 (C1S1)	41	0.88	0.29	1 (C1S1)
17	0.50	0.62	1 (C1S1)	42	0.89	0.62	1 (C2S1)
18	0.50	0.73	1 (C2S1)	43	0.89	0.73	2 (C2S2)
19	0.50	0.73	1 (C1S2)	44	0.90	0.73	2 (C1S3)
20	0.50	0.65	1 (C1S1)	45	0.90	0.73	2 (C2S2)
21	0.50	0.60	1 (C1S1)	46	0.91	0.77	3 (C3S2)
22	0.50	0.73	1 (C2S1)	47	0.92	0.62	1 (C2S1)
23	0.50	0.73	1 (C2S1)	48	0.93	0.73	2 (C2S2)
24	0.50	0.88	3 (C3S2)	49	0.96	0.73	1 (C2S1)
25	0.56	0.79	1 (C1S2)	50	0.99	0.82	3 (C2S3)

a The drug-effect relation can be 0 *not related*, 1 *doubtful*, 2 *possible*, 3 *probable* and 4 *highly likely*.

C = chronological criteria (C1 [*dubious*], C2 [*possible*] or C3 [*likely*]); **S** = semiological criteria (S1 [*dubious*], S2 [*possible*] or S3 [*likely*]).

i.e. *not related* and *highly likely,* have never been encountered.

When the French algorithm was compared with consensual expert judgement (table V) on the basis of the study sample, the *doubtful* score (37 cases) on the algorithm corresponded to an average probability of 0.52 (SD 0.24) with a range of 0.08–0.96, the *possible* score (eight cases) corresponded to a probability of 0.81 (range 0.40–0.93) and the *probable* score (five cases) to a probability of 0.81 (range 0.50–0.99).

For the 50 cases, the Spearman's correlation coefficient for probabilities from the logistic method and consensual expert judgement and categorized as defined in the Methods section was 0.439 ($p = 0.001$). The Spearman's correlation coefficient between the French algorithm and consensual expert judgement was 0.419 ($p = 0.002$). Sensitivity, specificity and positive and negative predictive values for both the logistic method and the algorithm are given in table VI. Values were

Table V. Distribution of probabilities of drug causality given by consensual expert judgement and the logistic method according to algorithm score

French algorithm intrinsic score	Experts' probability		Logistic probability	
	median	mean	median	mean
1 *(doubtful)*	0.50	0.52	0.62	0.63
2 *(possible)*	0.87	0.81	0.73	0.78
3 *(probable)*	0.85	0.81	0.82	0.81
Total	0.58	0.60	0.73	0.67

Table VI. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the logistic method and the French algorithm taking experts' consensus as reference, excluding neutral consensual expert judgements, i.e. p = 0.45–0.55

Parameter	Logistic method	French algorithm
Sensitivity	0.96	0.42
Specificity	0.42	0.92
PPV	0.78	0.92
NPV	0.83	0.42

computed on the basis of 12 'true negatives' and 26 'true positives' by excluding neutral experts' judgements, i.e. p = 0.45–0.55. Analysis of sensitivity and specificity on the whole sample, i.e. including the 12 non-discriminant assessments, gives results not significantly altered for both the logistic method and the French algorithm (table VII).

Discussion

Causality assessment of ADRs is like the 'quest for the holy grail', since a unique and operational tool providing an indisputable gold standard does not exist,[16,17] even if consensual expert judgement is generally considered as a reference.[18,19] One of the strengths of the present study is that it compares two operational approaches, i.e. French algorithmic and probabilistic, versus a valid reference on the basis of actually reported ADR cases, at least in the framework of routine pharmacovigilance and clinical practice. Moreover, the selected algorithm is, with that of Naranjo et al.,[20] one of the very few currently used on a large scale in routine pharmacovigilance.[14]

While a random sample of 50 spontaneous reports cannot encompass all the possible aspects of causality assessment, it presents two major advantages: (i) being randomly sampled, it reflects the reality of routine pharmacovigilance; and (ii) it avoids comparisons based on theoretical and/or seldom encountered situations such as cases with a very high or very low probability for drug causation. Indeed, the added value of causality assessment tools is clearly greater for decision making under uncertainty than for clearcut situations.

The 50 experts' assessments effectively illustrated this over-representation of uncertainty with a median of 0.58, a mean of 0.60 and 54% (n = 27) of cases included in the 0.25–0.75 range. This predominance is found in most sets of actual ADR cases, as well as in spontaneous reporting databases.[17,21,22]

Compared with the experts, the probability distribution obtained from the logistic method was skewed to the right with a median of 0.73, a mean of 0.67 and only six cases having a probability of drug causation of <0.50. More striking was the cluster centered around p = 0.73 (15 cases). This is the probability obtained for all case reports in which the temporal relationship between drug treatment sequences and events was considered as compatible and the semiological investigations did not allow a non-drug alternative cause to be ruled out.

Conversely, the discrete five-degree distribution obtained from the French algorithm was clearly skewed to the left, with the *doubtful* score accounting for 37 of the 50 cases (74%), even though for experts the probability for these 37 cases ranged from 0.08 to 0.96 with a median of 0.50 and a mean of 0.52, and 15 of these were associated with a probability of drug-causation of >0.50. As shown in table II, the *doubtful* score resulted from a temporal relationship between drug treatment and the event assessed as poor (19 cases), and/or from the absence of a typical clinical or biological pattern or from an alternative cause that was not ruled out (34 cases). This *doubtful* score seems more prevalent with the French algorithm than with other algorithms currently used (e.g. the Naranjo method),[23,24]

Table VII. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the logistic method and the French algorithm taking experts' consensus as reference, where neutral consensual expert judgements, i.e. p = 0.45–0.55, were considered as possibly in favour of drug responsibility

Parameter	Logistic method	French algorithm
Sensitivity	0.95	0.32
Specificity	0.42	0.92
PPV	0.84	0.92
NPV	0.71	0.30

and a high proportion of *doubtful* scores obtained using this method has already been reported, ranging from 68% to 92% according to studies.[14,22-24] In the paper by Thiessard et al.,[14] describing 197 580 spontaneous ADR reports to the French pharmacovigilance system from 1986 to 2001, causality assessment was *highly likely* for only 1% of the drugs, *probable* for 8%, *possible* for 17%, *doubtful* for 74% and *not related* for 1%. Interestingly, the proportion of *doubtful* assessments in this large series is exactly the same as in the present study sample.

It must be emphasized that individual expert assessments were not uniform in the initial phase of adjudication. In only 13 of 50 cases were all three judges in agreement within a 25% increment of the probability scale. Therefore, it is the averaged and/or consensual estimate that has been used as a gold standard in the assessment of the algorithmic and logistic methods.

As shown in table IV, the agreement of the logistic method with consensual expert judgement was good, except for the lower part of the probability scale, with seven major disagreements (case numbers 4, 6, 8, 9, 10, 11, 12). All correspond to a clear over-estimation of the probability given by the logistic method with a reversed conclusion about the plausibility of drug causation. On the other hand, the fit in the upper range of probabilities was acceptable, except for case number 41 where there are two clearcut opposite estimates. These results led to a moderate but significant correlation ($r = 0.439$; $p = 0.001$) between the logistic method and the consensual expert judgement. In their original article, Arimone et al.[13] found, as expected, a much higher correlation ($r = 0.86$), since the set of cases was the same as that used for weighting of the logistic method.

The agreement of the French algorithm with consensual expert judgement was acceptable for the lower range of probabilities since, with the exception of case numbers 12 and 24, 22 of 24 cases (92%) for which the probability of drug causation was ≤0.50 as assessed by the experts were assessed as *doubtful* (score 1) with the algorithm. Note that the lowest score, i.e. *not related*, was never encountered in the study sample because it was unlikely that a health professional reported an ADR case with evidence of non-drug responsibility. On the other hand, the agreement between the algorithm and consensual expert judgement was poor for the upper range of probabilities: for the 26 cases having a probability of drug causation of >0.50 for experts, i.e. 0.56–0.99, 15 were classified as *doubtful* with the algorithm. These results led to a moderate but significant concordance ($r = 0.419$; $p = 0.002$) between the algorithm and consensual expert judgement.

In summary, the logistic and algorithmic methods performed differently: the former tended to inflate the upper range and the latter to minimize it, which led to an overall poor agreement between these two procedures. The discrepancies were to some extent expected since, although considering roughly the same criteria, the two methods rely on different principles, e.g. the logistic method tries to mimic consensual expert judgement and to respect the probability rules. Another major difference is the weight associated with each criterion, with the French algorithm tending to highly increase the causality score when some criteria are assessed as positive (e.g. a positive rechallenge or a positive laboratory test). This somewhat excessive weighting of a positive rechallenge contrasts with the relatively low weighting of this event in the logistic method, considered as 'the cherry on the cake' of a previously both positive challenge and dechallenge. To a lesser extent, the expression of the final causality score as a 0–1 probability with the logistic method and in five discrete degrees with the algorithm could also play a part in the difference of results.

Regarding diagnostic tests, positive and negative predictive values are criteria of choice to evaluate the practical interest of a causality assessment tool. This is particularly true in pharmacovigilance, where both a high sensitivity and a high positive predictive value are to be preferred. Table VI shows the sensitivity, specificity and positive and negative predictive values, using consensual expert judgement as a reference, and after removing the 12 cases for which the causal relationship was assessed as neutral by

experts, i.e. having a probability ranging from 0.45 to 0.55 on the VAS.

As an illustration of their opposing performance, the logistic method demonstrated a high sensitivity (0.96) and low specificity (0.42), while the algorithm had low sensitivity (0.42) and high specificity (0.92). Interestingly, both positive and negative predictive values were good for the logistic method, 0.78 and 0.83, respectively, versus 0.92 and 0.42 for the algorithm. The diagnostic value for the French algorithm previously reported by Benahmed et al.[23] and Koh et al.[25] was similar to that reported here, with a low sensitivity (0.08 and 0.05, respectively) and a high specificity (98.3% and 100%, respectively). No concordance was found in the literature between the French algorithm and the Naranjo[23,24] or Jones' algorithms.[23] Therefore, although far from perfect, the logistic approach appears to be a more reliable procedure for routine assessment of ADR cases, as the probability of misclassification is lower with the logistic approach, particularly when concluding non-drug causation.

In view of its satisfactory sensitivity and positive predictive value, the logistic method could be applied to automated signal generation in ADR databases. Indeed, for the drug-event pairs judged positive (probability >0.55), the agreement between consensual experts' judgement and the logistic method was very high, with the exception of case number 41. These results suggest that the logistic method is able to identify all true signals in continuous surveillance; a subsequent validation of these by experts would allow the retention of the most relevant ones. This could be of great interest in routine pharmacovigilance. Nevertheless, it should be emphasized that with a substantial rate of false positives relative to true negatives (low specificity), the logistic method does not replace, but can be complemented by, the critical clinical assessment of individual cases in evaluating drug-related risk.

Another relevant finding of the present study, although based on a limited number of cases (n = 20), is that in the event of agreement between the logistic and algorithm procedures, the fit with consensual expert judgement was very close, with the exception of case numbers 12, 24 and 41.

One obvious limitation of the present study is the limited size of the study sample (i.e. 50) and case numbers of specific types of ADR. There may be important differences in performance characteristics of these methods based on different categories that have not been evaluated in this study.

With other studies using experts as a reference, the alternative option would have been to increase the number of cases at the expense of a less indisputable reference. Indeed, the gold standard was reliable because of the step-by-step consensual assessment of each case by a group of three senior experts who were able to refer to four other experts in the event of disagreement. However, such a validation process is not realistic when the number of cases is high.

Conclusions

To our knowledge, this study is the first to compare a probabilistic and an algorithmic drug causality assessment procedure (i) on the basis of actual ADR reports, and (ii) by using a consensus of senior experts.

The agreement between the two procedures was poor, with the probabilistic approach proving its superiority over the compared algorithm, which is widely used by the French pharmacovigilance network. The high sensitivity and positive predictive value of the probabilistic approach make it particularly interesting for routine assessment and automated screening of ADR signals in pharmacovigilance databases.

Acknowledgements

We thank Philip Robinson who kindly supervised the writing of this paper in English. This study was funded as a research project by a grant from the non-profit association ARME-Pharmacovigilance (Bordeaux, France). The authors have no conflicts of interest that are directly relevant to the content of this study.

References

1. Agbabiaka TB, Savovic J, Ernst E. Methods for causality assessment of adverse drug reactions: a systematic review. Drug Saf 2008; 31 (1): 21-37

2. Meyboom RH, Hekster YA, Egberts AC, et al. Causal or casual? The role of causality assessment in pharmacovigilance. Drug Saf 1997; 17: 374-89
3. Stephens MD. The diagnosis of adverse medical events associated with drug treatment. Adverse Drug React Acute Poisoning Rev 1987; 6: 1-35
4. Blanc S, Leuenberger P, Berger JP, et al. Judgments of trained observers on adverse drug reactions. Clin Pharmacol Ther 1979; 25: 493-8
5. Karch FE, Smith CL, Kerzner B, et al. Adverse drug reactions: a matter of opinion. Clin Pharmacol Ther 1976; 19: 489-92
6. Koch-Weser J, Sellers EM, Zacest R. The ambiguity of adverse drug reactions. Eur J Clin Pharmacol 1977; 11: 75-8
7. Kramer MS. Difficulties in assessing the adverse effects of drugs. Br J Clin Pharmacol 1981; 11 Suppl. 1: 105-10S
8. Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. Int J Forecast 1999; 15: 353-75
9. Lanctot KL, Naranjo CA. Comparison of the Bayesian approach and a simple algorithm for assessment of adverse drug events. Clin Pharmacol Ther 1995; 58: 692-8
10. Auriche M. Bayesian approach to the imputability of undesirable phenomena to drugs. Therapie 1985; 40: 301-6
11. Kramer MS. Imputabilité des effets indésirables: individu (analyse du cas) versus groupe (épidémiologie). 3es entretiens. Lyon: Jacques Cartier, 1989: 31-44
12. Péré JC, Godin MH, Bégaud B, et al. Sensitivity and specificity of imputability criteria: study and comparison of these efficacity indices for 7 methods. Therapie 1985; 40: 307-12
13. Arimone Y, Bégaud B, Miremont-Salamé G, et al. A new method for assessing drug causation provided agreement with experts' judgment. J Clin Epidemiol 2006; 59: 308-14
14. Thiessard F, Roux E, Miremont-Salamé G, et al. Trends in spontaneous adverse drug reaction reports to the French pharmacovigilance system (1986-2001). Drug Saf 2005; 28: 731-40
15. Bégaud B, Evreux JC, Jouglard J, et al. Imputation of the unexpected or toxic effects of drugs: actualization of the method used in France. Therapie 1985; 40: 111-8
16. Meyboom RH, Hekster YA, Egberts AC, et al. Causal or casual? The role of causality assessment in pharmacovigilance. Drug Saf 1997; 17: 374-89
17. Meyboom RH. Causality assessment revisited. Pharmacoepidemiol Drug Saf 1998; 7 Suppl. 1: S63-5
18. Karch FE, Lasagna L. Toward the operational identification of adverse drug reactions. Clin Pharmacol Ther 1977; 21: 247-54
19. Macedo AF, Marques FB, Ribeiro CF. Can decisional algorithms replace global introspection in the individual causality assessment of spontaneously reported ADRs? Drug Saf 2006; 29: 697-702
20. Naranjo CA, Busto U, Sellers EM, et al. A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 1981; 30: 239-45
21. Macedo AF, Marques FB, Ribeiro CF, et al. Causality assessment of adverse drug reactions: comparison of the results obtained from published decisional algorithms and from the evaluations of an expert panel, according to different levels of imputability. J Clin Pharm Ther 2003; 28: 137-43
22. Miremont G, Haramburu F, Bégaud B, et al. Adverse drug reactions: physicians' opinions versus a causality assessment method. Eur J Clin Pharmacol 1994; 46: 285-9
23. Benahmed S, Picot MC, Hillaire-Buys D, et al. Comparison of pharmacovigilance algorithms in drug hypersensitivity reactions. Eur J Clin Pharmacol 2005; 61 (7): 537-41
24. Eiden C, Peyrière H. Comparaison de deux méthodes d'imputabilité des effets indésirables du voriconazole notifiés dans la base nationale de pharmacovigilance: Begaud versus Naranjo. Pharmacien Hospitalier 2009; 44 (4): 186-9
25. Koh Y, Yap CW, Li SC. A quantitative approach of using genetic algorithm in designing a probability scoring system of an adverse drug reaction assessment system. Int J Med Inform 2008; 77 (6): 421-30

Correspondence: *Hélène Théophile*, Département de Pharmacologie, Université Victor Segalen, 33076 Bordeaux, Cedex, France.
E-mail: helene.theophile@pharmaco.u-bordeaux2.fr