Drug Saf (2013) 36:1033–1044
DOI 10.1007/s40264-013-0083-1

ORIGINAL RESEARCH ARTICLE

Comparison of Three Methods (An Updated Logistic Probabilistic Method, the Naranjo and Liverpool Algorithms) for the Evaluation of Routine Pharmacovigilance Case Reports Using Consensual Expert Judgement as Reference

Hélène Théophile · Manon André · Ghada Miremont-Salamé · Yannick Arimone · Bernard Bégaud

Published online: 5 July 2013

Abstract

Background An updated probabilistic causality assessment method and the Liverpool algorithm presented as an improved version of the Naranjo algorithm, one of the most used and accepted causality assessment methods, have recently been proposed.

Objective In order to test the validity of the probabilistic method in routine pharmacovigilance, results provided by the Naranjo and Liverpool algorithms, as well as the updated probabilistic method, were each compared with a consensual expert judgement taken as reference.

Methods A sample of 59 drug–event pairs randomly sampled from spontaneous reports to the French pharmacovigilance system was assessed by expert judgement until reaching consensus and by members of a pharmacovigilance unit using the updated probabilistic method, the Naranjo and Liverpool algorithms. Probabilities given by the probabilistic method, and categories obtained by both the Naranjo and the Liverpool algorithms were compared as well as their sensitivity, specificity, positive and negative predictive values.

Results The median probability for drug causation given by the consensual expert judgement was 0.70 (inter-quartile range, IQR 0.54–0.84) versus 0.77 (IQR 0.54–0.91) for the probabilistic method. For the Naranjo algorithm, the 'possible' causality category was predominant (61 %), followed by 'probable' (35 %), 'doubtful', and 'almost certain' categories (2 % each). Category distribution obtained with the Liverpool algorithm was similar to that obtained by the Naranjo algorithm with a majority of 'possible' (61 %) and 'probable' (30 %) followed by 'definite' (7 %) and 'unlikely' (2 %). For the probabilistic method, sensitivity, specificity, positive and negative predictive values were 0.96, 0.56, 0.92 and 0.71, respectively. For the Naranjo algorithm, depending on whether the 'possible' category was considered in favour or in disfavour of drug causation, sensitivity was, respectively, 1 or 0.42, specificity 0.11 or 0.89, negative predictive value 1 or 0.22 and positive predictive value 0.86 or 0.95; results were identical for the Liverpool algorithm.

Conclusion The logistic probabilistic method gave results closer to the consensual expert judgment than either the Naranjo or Liverpool algorithms whose performance were strongly dependent on the meaning given to the 'possible' category. Owing to its good sensitivity and positive predictive value and by providing results as continuous probabilities, the probabilistic method seems worthy to use for a trustable assessment of adverse drug reactions in routine practice.

H. Théophile · G. Miremont-Salamé · Y. Arimone · B. Bégaud
Univ de Bordeaux, U657, 33000 Bordeaux, France

H. Théophile · M. André · G. Miremont-Salamé · Y. Arimone · B. Bégaud
INSERM U657, 33000 Bordeaux, France

H. Théophile · G. Miremont-Salamé · B. Bégaud
Département de Pharmacologie, Centre Régional de Pharmacovigilance, CHU Bordeaux, 33000 Bordeaux, France

H. Théophile · Y. Arimone
Arme-P, 33000 Bordeaux, France

H. Théophile (✉)
Centre de Pharmacovigilance, Hôpital Pellegrin, CHU, 33076 Bordeaux Cedex, France
e-mail: helene.theophile@pharmaco.u-bordeaux2.fr

1 Background

Even if modern approaches developed during the last decades (e.g., automated data mining) strengthen the

efficiency of spontaneous reporting, the assessment of a possible causal link between a drug treatment and the occurrence of an adverse event in a given patient remains a complementary and valuable approach. The latter can be valuable for sorting out reports according to their degree of belief when statistical methods cannot discriminate between background noise and attributable cases, or for managing adverse events at the individual level to point at the most suspect drug or those that should be discontinued. This assessment roughly ensues from three main approaches: expert judgement or global introspection, algorithms, and probabilistic methods [1–3]. Clinician's judgement remains the first and indispensable step to identify and assess an adverse drug reaction, particularly if serious. Nevertheless, such an assessment, implicitly related to the diagnosis-making process, is by essence subjective and suffers from poor intra- and inter-rater reproducibility [4–7]. Causality assessment individually made by several senior experts and then interacting on a consensual basis (e.g., by using the Delphi method [8]) is, however, generally presented as a reference [9, 10], even if relatively complex to use in practice. Since the late 1970s, more than 20 different algorithms have been developed in an attempt to standardize the causality assessment process and to minimize the inter- and intra-rater variability. In addition to general algorithms that assess all adverse events and all drugs, some algorithms targeting an event (e.g., hepatitis) have been proposed. These algorithms include criteria specific to an event and allow a better assessment than non-specialized algorithms [11]. Although all these published algorithms are operational and relatively easy to use, none has been universally accepted [1]. Their weaknesses mainly ensue from the weighting arbitrarily assigned to each causality criterion by the authors of the method [12, 13] and from the sketchy expression of the final result, generally in the form of a x degree score. Conversely, probabilistic approaches, derived from Bayes' theorem, present the advantages of providing a formal causal assessment and giving results directly under the form of probabilities for drug causation. Although considered as the most reliable, these methods are, however, too complex and time consuming to be adopted in routine practice [1, 14].

In 2006, a novel causality assessment approach providing a probability of drug causation from assessment of seven causality criteria was proposed [15]. These assessments were statistically weighted by consensual expert judgement in order to directly provide a probability for drug causation by means of the logistic function: in the absence of any argument in favour or disfavour of the drug responsibility, the sum of weights is equal to 0 and the probability of drug causation fixed at 0.5; arguments in favour of drug responsibility make this sum of weights superior to 0 and the probability obtained ranges between 0.5 to 1; conversely, arguments in disfavour of drug responsibility make the sum of weights inferior to 0 and the probability less than 0.5. The probabilistic (logistic) method preserves the simplicity of algorithms, particularly when using the computerized version of the method that directly converts criteria assessment into probability for drug causation [16].

Since its first publication in 2006 [15], this method has been compared to a consensual expert judgement using a set of drug–event pairs representative of routine pharmacovigilance [17]. Its poor specificity (0.42) and a trend to overestimate drug responsibility led to an improvement in the logistic method [18]. A major change in this updated version (Table 1) concerned situations where investigations to rule out a possible alternative cause were required but not, or incompletely, performed; the final probability of drug causation being now decreased in such cases. Similarly, when investigations were not conducted since considered as not relevant, the weight assigned to its assessment is 0 and does not affect the probability of drug causation. The latter refers to situations where the event is, per se, evocative of drug aetiology (e.g., eruption occurring a few days after starting drug treatment, or bleeding with oral anticoagulants) or to symptoms or diseases for which no precise aetiology is known (e.g., multiple sclerosis).

One of the most widely used causality methods is the Naranjo algorithm, published in 1981 [19]. Although subject to some criticism [20–24], this algorithm, consisting of 10 simple questions, remains commonly used for the causality assessment of a suspected adverse drug reaction in spontaneous reports [25], case reports [26, 27] and observational studies [28, 29]. Recently, a new causality assessment tool, the Liverpool algorithm [30], has been proposed to overcome certain limits of the Naranjo algorithm.

In order to test the validity of the logistic method in routine pharmacovigilance (i.e., by using a study sample made of actual reports of various types), the Naranjo and Liverpool algorithms as well as the updated version of the logistic method were compared with consensual expert judgement that was used as the gold standard [10, 31].

2 Methods

The aim of our study was to evaluate the behaviour of each method when assessing a set of adverse events actually representative of daily pharmacovigilance. For this purpose, a sample of 59 drug–event pairs randomly sampled from spontaneous case reports to the French pharmacovigilance system was considered. Ten of these 59 drug–event pairs corresponded to particular situations seldom

Table 1 New version of the logistic method to obtain a probability of drug causation

	Statistical weights
Time to onset	
Incompatible	−5
Not suggestive	−0.48647
Unknown or not available	0
Compatible	+0.72218
Highly suggestive	+0.79190
Dechallenge	
Against the role of the drug	−1.32394
Non-conclusive or not available	0
Suggestive	+0.45961
Rechallenge	
Negative	−0.97045
Not attempted or not conclusive	0
Positive	+0.19114
Search for other aetiology	
Another cause highly probable	−2.74122
Required and not investigated or/and another possible cause	−1.04487
Not required and/or not applicable	0
Another cause ruled out	+0.16723
Risk factor(s) for drug reaction	
Ruled out or absent	0
Well validated and present	+1.18048
Reaction at site of application or plasma concentration known as toxic or validated laboratory test	
Unrelated or not available	0
Present or/and positive	+1.25352
Previous information on the drug and symptomatology	
Reaction not previously reported and type B	−0.42331
Not available	0
Not well known or previously published once or twice	+0.02686
Well known and labelled reaction	+0.36131

The sum of statistical weights is converted into probability by using the logistic function: $p = \frac{1}{1+e^{(-\sum \text{weights})}}$

met in routine (and thus unlikely to be found in a random sample) yet important to take into consideration to evaluate the performance of causality assessment methods. One drug–event pair was randomly sampled for each of these particular situations: drug overdose, drug–drug interaction, withdrawal syndrome, adverse event related to drug exposure during pregnancy, adverse event related to drug exposure during breastfeeding, adverse event at the site of drug application, adverse event with fatal outcome, adverse event with recurrence of signs when the suspected drug was rechallenged, adverse event with no recurrence of signs when the suspected drug was rechallenged, adverse event after vaccination (other than reaction at the injection site and effect occurring within 48 h after injection). For each case, information available from the complete file was summarized in standardized form, including the patient's characteristics, the suspected drug(s) with dates of treatment initiation and discontinuation, the type of event, its date of onset, relevant biological and clinical data, other current medicines and the time course of the event. The likelihood for drug causation for the 59 drug–event pairs was then assessed separately by two groups of experts.

2.1 Evaluation of Drug Causation by Consensual Expert Judgement

This approach was used as the gold standard. Each drug–event pair was assessed by a multidisciplinary group with expertise both in clinical pathology and in pharmacovigilance. For 31 cases, this group comprised three senior physicians who were heads of a pharmacovigilance regional centre or of a pharmacovigilance unit in a pharmaceutical company or of a department of internal medicine in a university hospital. For the remaining drug–events pairs ($n = 28$) that appeared to be more complex to assess (effect involving different organs, or different possible pathophysiological mechanisms, multiple risk factors, etc.) two additional experts were added. Each expert was asked to express separately his/her judgment on the responsibility of the suspected drug on a 100 mm visual analogue scale (VAS). The judgement was then directly converted into a probability of drug causation ranging from 0 to 1. Secondly, causes for discrepancies, defined by a difference of 25 mm on the VAS, were discussed by the same experts according to the Delphi process [8] and until agreement on a probability of drug responsibility. For each case, the final probability obtained by consensual agreement was retained as the gold standard for drug causation.

To comply with current practice of routine drug causality assessment, the pharmacovigilance medical staff of the Bordeaux pharmacovigilance centre (two senior clinical pharmacologists, one pharmacist and one physician with a large amount of experience in internal medicine) assessed the likelihood of drug causation for each of the 59 drug–event pairs, by first using the logistic method [16], followed by the Naranjo algorithm [19] one month later, and then the Liverpool algorithm [30].

2.2 Evaluation of Drug Causation by the Logistic Algorithm

The logistic algorithm [18] consists of assessing seven causality criteria: time to onset, dechallenge, rechallenge, search for other aetiology, risk factor(s) for drug reaction

(e.g., a drug–disease or drug–drug interaction increasing the toxicity of a drug), reaction at site of application and/or validated laboratory test clearly in favour of the drug responsibility, and previous reports or publication of similar drug–event associations and/or symptoms evocative of a drug causation; i.e., extrinsic plausibility (Table 1). The weights (β_1 to β_7) associated with the assessment of the seven criteria of the logistic method, symbolized by X_1, X_2,…,X_7, were obtained by using a multilinear regression model and converted into a probability p of drug causation by the logistic function:

$$p = \frac{1}{1 + \exp\left[-(\alpha + \sum_{i=1}^{7} \beta_i X_i)\right]}$$

α was fixed at 0 in order to obtain a 0.5 probability in the absence of argument in favour or disfavour of drug responsibility ($\sum \beta_i X_i = 0$).

2.3 Evaluation of Drug Causation by the Naranjo Algorithm

This algorithm [19] consists of assessing the causal relationship between a suspected drug and an adverse event by answering yes/no/do not know for ten successive questions: notoriety, temporal sequence, dechallenge, rechallenge, alternative causes, placebo response, drug levels in body fluids or tissues, dose-response relationship, previous history of similar adverse drug reaction and confirmation by objective evidence. The process results in a final score ranging from −4 to +13, allowing the qualification of drug responsibility by four categories: 'doubtful' (score ≤0), 'possible' (score between 1 and 4), 'probable' (score between 5 and 8), and 'almost certain' (score ≥9).

2.4 Evaluation of Drug Causation by the Liverpool Algorithm

This algorithm [30] or flowchart is based upon dichotomous questions that were issued from the Naranjo algorithm with some alterations: Question 6 on the placebo effect has been removed; Questions 7 and 10, both related to objective evidence of the adverse drug reaction occurrence, have become a single question; Question 8, related to the dose–effect relationship between drug and adverse event, is now combined with questions on time to onset and dechallenge. Moreover, two questions have been added: one on the flare-up of pre-existing symptoms by the drug and one on the temporal relationship for events resulting in long-lasting disability or impairment. Each answer is routed to a more specific question resulting in four causality categories: 'unlikely', 'possible', 'probable' and 'definite'.

2.5 Statistical Analysis

Descriptive analysis was based upon median, mean, range and quartiles for probabilities, i.e. for consensual expert judgement and logistic method, and upon distribution of causality categories for Naranjo and Liverpool algorithms. Correlation between the probabilities obtained from the logistic method and the consensual expert judgement on one hand, and the scores given by the Naranjo algorithm and the consensual expert judgement on the other hand, was assessed by the Spearman's correlation coefficient. Sensitivity, specificity, positive and negative predictive values of logistic, Naranjo and Liverpool methods were computed by using the consensual expert judgement as the gold standard. For expert judgement and the logistic method, probabilities lower than 0.50 were considered in disfavour of drug responsibility and those of 0.50 or more in favour of drug causation. For the Naranjo and Liverpool algorithms, two options were foreseen due to the non-explicit character of the 'possible' category: this category was first considered in favour of drug causation, as were 'probable' and 'almost certain' categories, and the 'doubtful' category in disfavour of drug responsibility; secondly, the 'possible' and 'doubtful' categories were considered in disfavour of drug responsibility while 'probable' and 'almost certain' categories were in favour of drug causation. Computations were made by using the STATA® software (version 8.2 for Macintosh, STATA Corporation, College Station, USA).

3 Results

3.1 Descriptive Analysis of Data

The description of the study sample is presented in Table 2. Out of the 59 drug–event pairs, 30 (51 %) were considered as serious (i.e., resulted in hospitalization or prolongation of hospitalization, disability/incapacity, life-threatening or death). The most often reported adverse drug reactions coded by organ system concerned the skin (20 %), gastrointestinal tract (17 %), nervous system (14 %), body as a whole or administration site conditions (10 %), followed by blood dyscrasias, cardiac, liver and respiratory injuries (7 % each). The most frequently involved drugs were antithrombotics and anti-infectives (17 % each), metabolism modifiers (14 %), cardiovascular treatments (10 %) and antineoplasics (7 %). There were two non-intentional and one intentional overdoses.

The initial individual assessment of the 59 drug–event pairs showed a relatively high degree of disagreement between experts: for only 21 cases (36 %) the judgement expressed by experts on the VAS was included in an

Table 2 Description of the random sample of the 59 drug–event pairs issued from the French pharmacovigilance database

Case no.	Gender	Age	Drug	Event(s)	Evolution	Known adverse drug reaction
1	M	48 y	Ursodésoxycholic acid	Liver fibrosis	Unknown	No
2	M	41 y	Paroxetine	Eruption	Recovered	Yes
3	M	16 y	Imatinib	Pneumothorax	Recovered	No
4	M	59 y	Rosiglitazone	Increased blood pressure, headache	Recovering	No
5	F	22 m	Cefatrizine	Urticaria	Recovered	Yes
6	F	6 m	Lamivudine	Developmental delay	Unassessable	No
7	M	30 y	Lenograstrim	Hepatitis	Recovered	Yes
8	M	71 y	Rimonabant	Nausea	Unknown	Yes
9	F	79 y	Urapidil	Vomiting, abdominal pain	Recovered	Yes
10	F	70 y	Fludarabine	Progressive multifocal leukoencephalopathy	Death	Yes
11	M	48 y	Parenteral nutrition	Liver fibrosis	Unknown	Yes
12	F	73 y	Insulin	Dysgeusia	Recovered	No
13	F	31 y	Oral bacterial vaccine[a]	Atopic dermatitis	Unassessable	Yes
14	M	57 y	lysine acetylsalicylate	Rectal bleeding	Recovered	Yes
15	M	58 y	Melphalan	Pancytopenia	Unknown	Yes
16	M	3 m	Azathioprine	Neutropenia, anaemia	Recovered	Yes
17	F	67 y	Fluindione	Epistaxis	Recovering	Yes
18	F	75 y	Levodopa/benserazide	Respiratory discomfort, rhinitis	Recovered	No
19	F	18 y	Human papillomavirus vaccine	Atopic dermatitis	Recovered	No
20	M	86 y	Clarithromycin	Hypoglycaemia[c]	Recovered	Yes
21	M	78 y	Buflomedil	Constipation	Recovering	No
22	M	35 y	Amoxicillin/clavulanic acid	Eruption	Recovered	Yes
23	F	71 y	Decongestant solution for inhalation[b]	Asthma, coma	Recovered	Yes
24	F	60 y	Salazopyrin	Erythema	Recovered	Yes
25	M	83 y	Ticlopidine	Cerebral haemorrhage	Death	Yes
26	F	77 y	Fondaparinux	Cerebral haemorrhage	Death	Yes
27	M	29 y	Vibramycin	Erythrodermia, phototoxicity	Recovered	Yes
28	F	85 y	Piroxicam	Erythema multiforme	Recovered	Yes
29	M	70 y	Valsartan	Photodistributed eczema	Recovering	No
30	F	91 y	Lisinopril	Decreased blood pressure, fall	Recovered	Yes
31	F	45 y	Amorolfine	Contact eczema	Recovered	Yes
32	F	2 y	Tixocortol	Urticaria	Recovered	Yes
33	M	67 y	Nilotinib	Increased QT	Recovered	Yes
34	M	55 y	Voriconazole	Torsade de pointes	Death	Yes
35	F	UK	Acarbose	Diarrhoea	Recovered	Yes
36	F	86 y	Diltiazem	Precordial pain, bradycardia[d]	Recovered	Yes
37	M	81 y	Nefopam	Mental confusion	Recovered	Yes
38	M	72 y	Amoxicillin/clavulanic acid	Diarrhoea	Recovered	Yes
39	F	86 y	Lysine acetylsalicylate	Ulcer, melena, anaemia	Unknown	Yes
40	F	86 y	Atenolol	Precordial pain, bradycardia[d]	Recovered	Yes
41	F	72 y	Enoxaparin	Digestive haemorrhage	Recovered	Yes

Table 2 continued

Case no.	Gender	Age	Drug	Event(s)	Evolution	Known adverse drug reaction
42	F	68 y	Digoxin	Malaise	Recovered	Yes
43	F	19 y	Pseudoephedrine	Meningeal syndrome: hypertension, headache, vomiting	Recovered	Yes
44	F	78 y	Fluindione	Rectal bleeding	Death	Yes
45	M	86 y	Glibenclamide	Hypoglycaemia[c]	Recovered	Yes
46	F	56 y	Hydroxyzine	Withdrawal syndrome	Recovered	Yes
47	M	96 y	Fluindione	Increased INR	Recovered	Yes
48	M	74 y	Fluindione	Peri-renal haemorrhage	Recovering	Yes
49	F	23 y	Ketoprofen	Erythema	Recovering	Yes
50	M	72 y	Iobitridol	Eruption, cough	Recovered	Yes
51	M	98 y	Acenocoumarol	Psoas haematoma, anaemia	Recovered	Yes
52	M	73 y	Gadoteridol	Quincke's edema	Recovered	Yes
53	F	82 y	Insulin	Injection site reaction	Recovering	Yes
54	M	67 y	Insulin	Hypoglycaemia	Recovered	Yes
55	M	77 y	Atorvastatin	Red urine	Recovered	No
56	M	5 y	B meningococcal vaccine	Injection site reaction	Recovered	Yes
57	F	40 y	Acetaminophen	Hepatitis, acute renal insufficiency	Recovered	Yes
58	M	6 d	Morphine	Hypotonia	Recovered	Yes
59	M	69 y	Iobitridol	Anaphylactic reaction	Recovered	Yes

d day, *INR* international normalized ratio, *m* month, *UK* unknown, *y* year

[a] Oral bacterial vaccine = ribosome fractions from haemophilus influenzae, *Klebsiella pneumoniae* membrane proteoglycans, *Klebsiella pneumoniae* ribosomal fraction, *Streptococcus pneumoniae*, *Streptococcus pyogenes*

[b] Inhalation solution = benzoin, eucalyptus tinctures, lavender and thyme extracts, L-menthol, peru balsam

[c] Pharmacokinetic interaction

[d] Pharmacodynamic interaction

interval of 25 mm. Reasons for disagreements were discussed during a meeting until a consensual probability of drug causation was reached. Causality assessments obtained on the study sample from the consensual experts judgement, the updated logistic method, the Naranjo and Liverpool algorithms are presented in Table 3.

3.2 Comparison of Causality Assessments Given by the Updated Logistic Method with those of the Consensual Expert Judgement

For the 59 studied drug–event pairs, the probabilities of drug causation obtained from the consensual expert judgement ranged from 0.01 to 0.98, with a mean of 0.68, a median of 0.70 and an inter-quartile range of 0.54–0.84 (Table 4). For the updated logistic method, probabilities ranged from 0 to 0.98 with a mean of 0.71, a median of 0.77 and an inter-quartile range of 0.54–0.91. A difference in probability between the updated logistic method and the consensual expert judgement greater than 0.25 was found in 6 of the 59 drug–event pairs (Cases 5, 13, 14, 16, 17, 55; i.e. 10.1 %) leading to a reversed conclusion about the plausibility of drug causation. For four of these six discordant drug–event pairs, assessments given by the consensual expert judgement were non-discriminant (probabilities of 0.51, 0.53, 0.55 and 0.55, respectively). Therefore, it can be considered that truly reversed conclusions between experts and the logistic method were only found in two cases (3.4 %). Other minor differences in probability ranging between 0.17 and 0.24 were observed but did not significantly alter the conclusion regarding drug responsibility (Cases 4, 24, 31, 42, 43 and 46). The distribution of probabilities obtained from the updated logistic method (Fig. 1) was significantly correlated with those of the consensual expert judgement (correlation coefficient 0.73, $p < 0.0001$). Considering a cut-off of 0.50 for

Table 3 Causality assessments given by the four approaches for the 59 drug–event pairs

Case no.	Expert probability	Logistic method probability	Naranjo category	Liverpool category	Case no.	Expert probability	Logistic method probability	Naranjo category	Liverpol category
1	0.01	0	Doubtful	Unlikely	31	0.7	0.84	Possible	Possible
2	0.15	0.23	Possible	Possible	32	0.72	0.83	Probable	Probable
3	0.18	0.32	Possible	Possible	33	0.75	0.62	Possible	Possible
4	0.25	0.43	Possible	Possible	34	0.75	0.62	Possible	Possible
5	0.30	0.78	Probable	Probable	35	0.75	0.77	Possible	Possible
6	0.47	0.43	Possible	Possible	36	0.77	0.84	Possible	Possible
7	0.47	0.62	Possible	Possible	37	0.78	0.62	Possible	Possible
8	0.48	0.51	Possible	Possible	38	0.78	0.85	Probable	Probable
9	0.48	0.51	Possible	Possible	39	0.80	0.77	Possible	Possible
10	0.5	0.51	Possible	Possible	40	0.8	0.85	Possible	Possible
11	0.5	0.51	Possible	Possible	41	0.81	0.84	Possible	Possible
12	0.51	0.43	Possible	Possible	42	0.81	0.98	Probable	Probable
13	0.51	0.78	Probable	Definite	43	0.83	0.62	Possible	Possible
14	0.53	0.85	Probable	Probable	44	0.84	0.91	Probable	Probable
15	0.54	0.51	Possible	Possible	45	0.84	0.94	Probable	Probable
16	0.55	0.82	Possible	Possible	46	0.86	0.62	Possible	Possible
17	0.55	0.94	Probable	Probable	47	0.87	0.94	Probable	Probable
18	0.57	0.54	Possible	Possible	48	0.88	0.94	Probable	Probable
19	0.58	0.54	Possible	Possible	49	0.88	0.94	Probable	Probable
20	0.61	0.54	Possible	Possible	50	0.92	0.95	Probable	Probable
21	0.61	0.64	Possible	Possible	51	0.93	0.83	Probable	Probable
22	0.63	0.62	Possible	Possible	52	0.93	0.94	Probable	Probable
23	0.63	0.64	Possible	Possible	53	0.93	0.96	Almost certain	Definite
24	0.63	0.82	Possible	Possible	54	0.94	0.94	Probable	Probable
25	0.65	0.51	Possible	Possible	55	0.95	0.48	Possible	Possible
26	0.67	0.75	Possible	Possible	56	0.95	0.98	Probable	Definite
27	0.68	0.77	Possible	Possible	57	0.96	0.95	Probable	Probable
28	0.69	0.54	Possible	Possible	58	0.97	0.95	Probable	Probable
29	0.69	0.62	Possible	Possible	59	0.98	0.98	Probable	Definite
30	0.70	0.91	Probable	Probable					

retaining drug responsibility, the updated logistic method, when compared with the consensual expert judgement, had a sensitivity of 0.96, a specificity of 0.56, positive and negative predictive values of 0.92 and 0.71, respectively, and an accuracy of 0.90 (Table 5).

3.3 Comparison of Causality Assessments given by the Naranjo Algorithm with those of the Consensual Expert Judgement

For the Naranjo algorithm, only 10 of the 18 possible causality scores (i.e., from -4 to $+13$) were met; the minimum (-4 and -3) and maximal (10 and over) scores were never observed for the 59 drug–event pairs assessed (Fig. 2). Considering the four causality categories obtained from scores, as indicated in Table 4, 'possible' ($n = 36$, 61 %) and 'probable' ($n = 21$, 35 %) categories were clearly predominant, whereas the two extremes, 'doubtful' and 'almost certain', were seldom encountered ($n = 1, 2$ % each).

Comparing the Naranjo algorithm with the consensual expert judgement with regards to the 59 drug–event pairs, the 'doubtful' category (a single case) from the Naranjo algorithm corresponded to a probability of drug causation of 0.01; the 'possible' category (36 cases) corresponded to a median probability of 0.63 (range 0.15–0.95); the 'probable' category (21 cases) to a median probability of 0.87 (range 0.30–0.98) and the 'almost certain' category (one single case) to a probability of 0.93 (Table 4). Results given by the Naranjo algorithm and the consensual expert judgement (Fig. 2) were significantly correlated (correlation coefficient 0.70, $p < 0.0001$). For the

Table 4 Distribution of probabilities of drug causality for the 59 drug–event pairs given by the consensual expert judgement and the new version of the logistic method according to the Naranjo algorithm

Naranjo scores	N	Consensual expert judgement						Updated logistic method					
		Mean ± standard deviation	Median	1st quartile	3rd quartile	Min	Max	Mean ± standard deviation	Median	1st quartile	3rd quartile	Min	Max
Doubtful (score ≤ 0)	1	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0	0	0
Possible (1 ≤ score ≤ 4)	36	0.62 ± 0.18	0.63	0.51	0.75	0.15	0.95	0.61 ± 0.15	0.62	0.51	0.75	0.23	0.85
Probable (5 ≤ score ≤ 8)	21	0.80 ± 0.18	0.87	0.72	0.93	0.30	0.98	0.91 ± 0.06	0.94	0.85	0.95	0.78	0.98
Almost certain (score ≥ 9)	1	0.93	0.93	0.93	0.93	0.93	0.93	0.96	0.96	0.96	0.96	0.96	0.96
Total	59	0.68 ± 0.22	0.70	0.54	0.84	0.01	0.98	0.71 ± 0.22	0.77	0.54	0.91	0	0.98

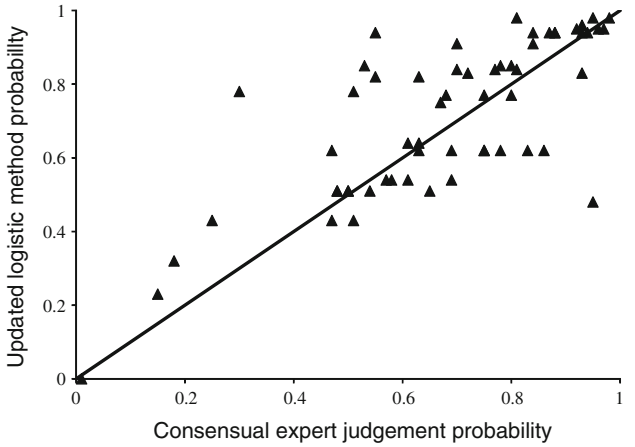

Fig. 1 Distribution of causality assessments given by the updated logistic method and by the consensual expert judgement for the 59 drug–event pairs (Spearman rank correlation coefficient = 0.73; $p < 0.0001$); *triangles are arranged on the line* in case of perfect agreement between the updated logistic method and the consensual expert judgement

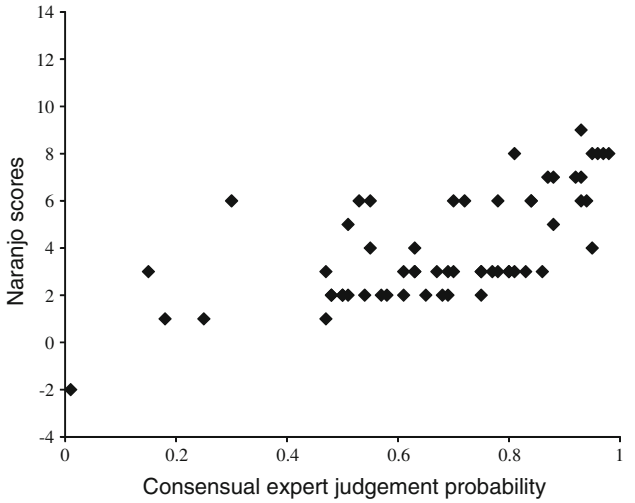

Fig. 2 Distribution of causality assessments given by the Naranjo method and the consensual expert judgement for the 59 drug–event pairs (Spearman rank correlation coefficient = 0.70; $p < 0.0001$)

assessment of the sensitivity, specificity, positive and negative predictive values of the Naranjo algorithm, the 'possible' category was successively considered in favour and in disfavour of drug causation. Results presented in Table 5 showed a sensitivity of 1, a specificity of 0.11 and a negative predictive value of 1 in the first case and a sensitivity of 0.42, a specificity of 0.89 and a negative predictive value of 0.22 in the second; the positive predictive value remaining not significantly altered (0.86 and 0.95, respectively). The accuracy of the Naranjo algorithm was better when the 'possible' category was considered in favour (0.86) rather than in disfavour (0.49) of drug causation.

3.4 Comparison of Causality Assessments Given by the Liverpool Algorithm Both to Those of the Consensual Expert Judgement and the Naranjo Algorithm

Of the 59 drug event pairs assessed with the Liverpool algorithm, the 'possible' category was predominant ($n = 36$, 61 %), followed by 'probable' ($n = 18$, 30 %), 'definite' ($n = 4$, 7 %) and 'unlikely' ($n = 1$, 2 %). The causality category distribution was quite similar to those of the Naranjo algorithm with a change in categories for only three drug–event pairs (Cases 13, 56 and 59); all corresponding to a shift of one causality degree from 'probable' to 'definite'. The assessment obtained by the Liverpool algorithm was closer to the consensual expert judgement for two of the three 'definite' drug–event pairs (Cases 56 and 59: probability of 0.95 and 0.98, respectively) and overestimated for one drug–event pair (Case 13 assessed at 0.51 by experts). As alterations made in the Liverpool algorithm concerned only 'probable' and 'certain' categories, both considered in favour of drug causation, sensitivity, specificity, negative and positive predictive values remained identical to those of the Naranjo algorithm (Table 5).

4 Discussion

In the present study, three causality assessment methods (a probabilistic approach, a widely used algorithm and its recently updated version) were compared with consensual expert judgement by using a random sample of actual drug–event pairs reported to a national pharmacovigilance system. For the logistic method, the distribution parameters obtained on the study sample were quite similar to those from the consensual expert judgement. The causes of discrepancies between both approaches mainly concerned (i) the assessment of the criterion 'search for other aetiology for the event', i.e., to consider or not another possible explanation as likely or if complementary investigations were required, and (ii) the somewhat elusive distinction between a risk factor for drug causation and another possible explanation. These disagreements, shared with other published algorithms and previously discussed [19, 32, 33], are likely to reflect real differences in clinical opinion. Beyond these points, the analysis of sensitivity and specificity of the logistic method showed that its performance was good overall when compared with the consensual expert judgement.

When compared with the consensual expert judgement, the Naranjo algorithm failed in the study sample to generate the highest and the lowest categories (i.e., 'doubtful' and 'almost certain') as previously found in other published studies [20, 22]. Garcia-Cortes et al. reported that for the 249 cases of suspected drug-induced liver injuries assessed by two clinicians by using the Naranjo algorithm, the percentages of 'doubtful' and 'almost certain' categories were, respectively, 2.2 and 0.2 % [20]. Similarly, Koh and Li [22], using the Naranjo algorithm for the assessment of 450 adverse drug reactions, found 0.2 % of 'doubtful' and 0.9 % of 'almost certain'. To increase the discriminative power of this algorithm, these authors suggested a cut-off score at 7 instead of 9 for the 'almost certain' category [23]. In the present study sample, the graphical analysis of the score distribution suggested that this cut-off at 7 would be more appropriate (Fig. 2). Actually, if causality assessments generated by the Naranjo algorithm focused on the 'possible' and 'probable' categories, one should note that on the basis of the 59 drug–event pairs, the 'possible' category corresponded for the experts to a very large range of probabilities (0.15–0.95). Teschke et al. [24], in their study on the assessment of herbal-induced hepatotoxicity, also noticed that the Naranjo algorithm concluded to 'possible' under almost any condition, even when

Table 5 Sensitivity, specificity, positive predictive value (*PPV*), negative predictive value (*NPV*) and accuracy of the updated logistic, the Naranjo and Liverpool methods, taking experts consensus as reference

	Logistic method	Naranjo method 1	Naranjo method 2	Liverpool method 1	Liverpool method 2
Sensitivity	0.96	1	0.42	1	0.42
Specificity	0.56	0.11	0.89	0.11	0.89
PPV	0.92	0.86	0.95	0.86	0.95
NPV	0.71	1	0.22	1	0.22
Accuracy	0.90	0.86	0.49	0.86	0.49

For the consensual expert judgement and the updated logistic method, probabilities <0.50 were considered in disfavour of drug responsibility and those ≥0.50 in favour of drug causation. For the Naranjo and the Liverpool methods, two options were taken into account for the possible score: in Naranjo and Liverpool methods 1, 'doubtful' or 'unlikely' score was considered in disfavour of drug responsibility and 'possible', 'probable' and 'almost certain' or 'definite' scores in favour of drug causation; in Naranjo and Liverpool methods 2, 'doubtful' or 'unlikely' and 'possible' scores were considered in disfavour of drug responsibility and 'probable' and 'almost certain' or 'definite' scores in favour of drug causation

there was lack of a clear temporal association or when the liver injury was not ascertained. This disrupting role of the 'possible' category is highlighted in the present study by the sensitivity analysis giving quite different results for sensitivity and specificity according to whether one considers that 'possible' suggests a drug causation or not. Analysis of the literature finds that the 'possible' category in the Naranjo algorithm is both viewed as in favour [10] or in disfavour [20, 34] of drug causation depending on the publications considered. Consequently, sensitivity and specificity of the Naranjo algorithm found in the literature are heterogeneous, with values ranging from 0.5 to 1 for sensitivity and from 0 to 1 for specificity [10, 20, 34]. However, the published positive predictive values are always high (between 0.81 and 0.95), which is concordant to that found in the present study.

The recently published Liverpool algorithm considered some of the criticisms previously raised for the Naranjo algorithm: (i) some questions per se non-explicit were clarified or removed [21]; (ii) the somewhat arbitrary weighting of the Naranjo causality criteria [21, 23] was resolved by a flowchart of dichotomous questions; and (iii) the 'definite' causality category seldom or never met in the original version [20, 21] was made more likely to achieve. In the study sample used here, the Liverpool algorithm showed a slightly increased discriminative power by classifying three more drug–event pairs in the 'definite' category. Nevertheless, disagreements with expert judgement persisted in certain situations: for example, the Liverpool algorithm did not ascribe the 'definite' category in cases of overdose (Cases 54, 57, 58) supported by a supra-therapeutic plasma drug level, error of administration, or dose-dependent events. In fact, by requiring a 'positive rechallenge' or a 'past history of the same event with the same drug' to obtain the 'definite' category, the Liverpool algorithm refers to situations that are rarely met in routine practice. Therefore, it was then not surprising that in the study sample composed of run-of-the-mill drug–event pairs reported to a national pharmacovigilance system, the Liverpool algorithm classified only three (5 %) more drug event pairs in the 'definite' category. These results differ from those of Gallagher et al. [20, 21], who observed in the three samples used for the development and the validation of the Liverpool algorithm an important increase (42, 13, and 46 %) of cases assessed as 'definite'. As suggested by Gallagher et al., this can be explained in two of the three samples by the presence of a large number of children with malignancies who had presented during repeated courses of chemotherapy a 'past history of the same event with the same drug'. In their third sample, Gallagher et al. analysed adverse drug reactions published in the journal *Annals of Pharmacotherapy*. Published case reports were likely to present with convincing arguments; e.g., extensive investigations of alternate aetiologies and positive rechallenge, both prone to favour the 'definite' category in the Liverpool algorithm. In practice, cases reported in the framework of routine pharmacovigilance are seldom so demonstrative [35] and most often present another possible aetiology to explain the adverse event; decision tools being, by essence, mainly useful in these situations. Moreover, the Liverpool algorithm by using a flow chart in the form of channelled questions could lead to a loss of some relevant information and to an underestimation of the drug causation. For example, if there was a low probability that the event was due to an underlying disease and objective evidence supportive of the causal adverse drug reaction mechanism, e.g. supra-therapeutic drug levels, this data is not taken into account in the Liverpool flow chart. Similarly, a 'positive rechallenge' in presence of a 'high (or unsure) probability that the event was due to an underlying disease' does not alter the causality score in the Liverpool algorithm. Even if, in the event of a positive rechallenge, the probability that the event could be due to an underlying disease was inevitably low, the Liverpool algorithm gives a very high weight to the rechallenge, whatever the other possible causes for the event.

Despite these points, both the Naranjo and the Liverpool algorithms present indisputable advantages thanks to their simplicity to use. Moreover, the final causality assessment, expressed in four qualitative categories, 'doubtful', 'possible', 'probable', and 'almost certain', is quite suitable for pharmacovigilance purpose as what matters is whether the drug was rather or not the cause of the event. The remaining problem concerns the cut-off between categories, arbitrarily fixed either by weighted criteria (Naranjo algorithm) or by a flow chart with arrowed questions (Liverpool algorithm). The logistic method overcomes this limit by generating continuing values from 0 to 1 directly interpretable under the form of a probability which has a priori the same meaning for everyone. The assessment process is less basic, since it involves considering several options for the assessment of each criterion, in particular for the search for alternative causes, but is made easier thanks to the computerized version of the method that offers numerous explanations and helps at each assessment step [16].

4.1 Limitations

In the current study, 24 % of assessments given by the consensual expert judgement had a probability of drug causation above 0.85. This is, at least in part, related to the random character of the study sample that included 10 cases of bleeding under antithrombotics, an adverse drug reaction clearly related to the pharmacological properties of the drug, as well as to several well described events such

as three drug overdoses. One could suspect the study sample not to be representative of adverse drug reactions reported to pharmacovigilance systems. However, in their analysis of adverse drug reactions reported to the French pharmacovigilance between 1986 and 2001, Thiessard et al. [36] showed that bleeding disorders with vitamin K antagonists were among the most frequent reported adverse drug reactions. Moreover, the most prevalent therapeutic classes (anti-infectives, metabolism and cardiovascular agents) and system organs of adverse drug reactions (skin, gastro-intestinal disorders, central nervous system) reported to the French pharmacovigilance were also the most prevalent in the present study sample [36]. Moreover, with consensual expert assessments that covered all the possible range of probabilities from 0 to 0.98, the studied sample included various clinical situations, which is of most importance when one intends to compare and test the validity of causality assessment methods.

The limited size of the sample could also be considered as a limitation of the present study. Clearly such a sample was far to test and compare the three methods by exploring all possible situations, i.e. types of drugs and type of effects, met in practice. The alternative option would have been to dramatically increase the number of cases at the expense of the validity of the reference. Indeed, the gold standard used in the present case was trustable only thanks to the step-by-step consensual assessment obtained for each case from a group of senior experts. Such a time-consuming validation process would not be realistic for a markedly larger number of cases. Another alternative would have been to include with the current set of cases the 50 drug–event pairs that had previously been used for the comparison of the logistic method and the French algorithm with the consensual expert judgement [17]. Unfortunately, it was not acceptable since the new weighting of the logistic method [18] was made on this prior set of 50 drug–event pairs.

The current sample, although of limited size, provided results consistent with those of previous studies that included many more cases [10, 20]. Garcia-Cortes et al. [20]. evaluated the accuracy of the Naranjo algorithm on 225 cases of suspected hepatotoxicity by considering the 'possible' category in disfavour of drug causation. They found moderate sensitivity (0.54) and good specificity (0.88), which is quite close to our results. Macedo et al. [10] assessed 500 adverse drug reports by considering the 'possible' category of the Naranjo algorithm in favour of drug causation. They found a high sensitivity (1) and a poor specificity (0) as in the current study.

Finally, the last point to be addressed concerns the validity of the reference used in the present study. Sadly, as in many other domains of medicine or human sciences, an undisputable gold standard for drug causation, such as an experimental proof, does not exist. The most sensible approach was not relying upon a single expert assessment but upon the two-step Delphi process described in the Method section. However, it could be argued that in some cases the consensual agreement could have been obtained on an erroneous point estimate. Even if theoretical, this possible limitation should be kept in mind.

5 Conclusions

The logistic probabilistic method provided results that were closer to those of consensual expert judgment than either the Naranjo or Liverpool algorithms, whose performance were strongly dependent on the 'possible' category, which includes a wide variety of probabilities of drug causation. By providing results expressed as continuous probabilities, the logistic method provides results directly interpretable and overcomes the problem of a more or less arbitrary cut-off between categories. Another advantage of the logistic approach is its good sensitivity and positive predictive value, two parameters to be favoured in the pharmacovigilance setting.

Acknowledgments The authors wish to thank to Philip Robinson (Univ of Bordeaux) for his advice and help in writing this paper. No sources of funding were used to assist in the preparation of the manuscript. Hélène Théophile, Manon André, Ghada Miremont-Salamé, Yannick Arimone and Bernard Bégaud declare no conflicts of interest that could have directly or indirectly affected the content of this manuscript.

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):105S–10S.
8. Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. Int J Forecast. 1999;15:353–75.
9. Karch FE, Lasagna L. Toward the operational identification of adverse drug reactions. Clin Pharmacol Ther. 1977;21:247–54.
10. 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.
11. Teschke R, Schulze J. Suspected herbal hepatotoxicity: requirements for appropriate causality assessment by the US Pharmacopeia. Drug Saf. 2012;35(12):1091–7.
12. Auriche M. Bayesian approach to the imputability of undesirable phenomena to drugs. Therapie. 1985;40:301–6.
13. Pere JC, Begaud B, Haramburu F, et al. Computerized comparison of six adverse drug reaction assessment procedures. Clin Pharmacol Ther. 1986;40(4):451–61.
14. Lanctot KL, Naranjo CA. Computer-assisted evaluation of adverse events using a Bayesian approach. J Clin Pharmacol. 1994;34(2):142–7.
15. 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.
16. Théophile H, Arimone Y, André M, et al. Computerized version of the logistic method. http://www.arme.u-bordeaux2.fr/. Accessed 7 May 2013.
17. Theophile H, Arimone Y, Miremont-Salame G, et al. 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. Drug Saf. 2010;33(11):1045–54.
18. Theophile H, André M, Arimone Y, et al. An updated method improved the assessment of adverse drug reaction in routine pharmacovigilance. J Clin Epidemiol. 2012;65(10):1069–77.
19. 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.
20. Garcia-Cortes M, Lucena MI, Pachkoria K, et al. Evaluation of Naranjo adverse drug reactions probability scale in causality assessment of drug-induced liver injury. Aliment Pharmacol Ther. 2008;27(9):780–9.
21. Kane-Gill SL, Kirisci L, Pathak DS. Are the Naranjo criteria reliable and valid for determination of adverse drug reactions in the intensive care unit? Ann Pharmacother. 2005;39(11):1823–7.
22. Koh Y, Li SC. A new algorithm to identify the causality of adverse drug reactions. Drug Saf. 2005;28(12):1159–61.
23. 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.
24. Teschke R, Schmidt-Taenzer W, Wolff A. Spontaneous reports of assumed herbal hepatotoxicity by black cohosh: is the liver-unspecific Naranjo scale precise enough to ascertain causality? Pharmacoepidemiol Drug Safety. 2011;20(6):567–82.
25. FDA. Drugcite: drug side effects reported to the FDA. http://www.drugcitecom/. Accessed 7 May 2013.
26. Instructions to Authors for submission case reports in Pharmacotherapy. www.pharmacotherapy.org/. Accessed 7 May 2013.
27. Instructions to Authors for submission case reports in The Annals of Pharmacotherapy. www.theannals.com/. Accessed 7 May 2013.
28. Davies EC, Green CF, Taylor S, et al. Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. PLoS One. 2009;4(2):e4439.
29. Reis AM, Cassiani SH. Adverse drug events in an intensive care unit of a university hospital. Eur J Clin Pharmacol. 2011;67(6):625–32.
30. Gallagher RM, Kirkham JJ, Mason JR, et al. Development and inter-rater reliability of the Liverpool adverse drug reaction causality assessment tool. PLoS One. 2011;6(12):e28096.
31. Karch FE, Lasagna L. Toward the operational identification of adverse drug reactions. Clin Pharmacol Ther. 1977;21(3):247–54.
32. Hutchinson TA, Leventhal JM, Kramer MS, et al. An algorithm for the operational assessment of adverse drug reactions. II. Demonstration of reproducibility and validity. JAMA. 1979;242(7):633–8.
33. Leventhal JM, Hutchinson TA, Kramer MS, et al. An algorithm for the operational assessment of adverse drug reactions. III. Results of tests among clinicians. JAMA. 1979;242:1991–4.
34. Miljkovic MM, Dobric S, Dragojevic-Simic V. Consistency between causality assessments obtained with two scales and their agreement with clinical judgments in hepatotoxicity. Pharmacoepidemiol Drug Saf. 2011;20(3):272–85.
35. Theophile H, Schaeverbeke T, Miremont-Salame G, et al. Sources of information on lymphoma associated with anti-tumour necrosis factor agents: comparison of published case reports and cases reported to the French pharmacovigilance system. Drug Saf. 2011;34(7):577–85.
36. 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.