Eur Arch Psychiatr Neurol Sci (1985) 235:158–163

European Archives of Psychiatry and Neurological Sciences

Reliability of Adverse Drug Reaction Assessment in Psychiatric Inpatients*

R. Grohmann[1], P. Dirschedl[3], J. Scherer[1], L. G. Schmidt[2], and O. Wunderlich[1]

[1] Psychiatrische Universitätsklinik, Nußbaumstrasse 7, D-8000 München 2,
[2] Psychiatrische Klinik der Freien Universität Berlin, D-1000 Berlin
[3] Abteilung für medizinische Statistik und Datenverarbeitung der Universität München, D-8000 München, Federal Republic of Germany

Summary. Within an ongoing drug surveillance program in psychiatric hospitals the applicability of an algorithm for judgment on probability of causal relationship of adverse events and drug therapy was tested. Algorithmic interrater agreement was compared to agreement obtained with the conventional criteria used so far within the program in 80 cases by two raters, who had participated in the drug surveillance program since its beginning in 1979. With the use of the algorithm raters agreed on imputed drugs in 86% of all cases; total agreement on drugs and degree of probability was obtained in 69% (weighted kappa 0.618). Raters agreed on total score for the imputed medication in 49% and also on all subscores for the different axes of the algorithm in 43% of all cases.

Differentiation of drug-related from illness-related changes, the use of judgmental terms within the algorithm and specific problems created by the frequent use of combinations of drugs with similar profiles of adverse drug reactions (ADR) in psychiatric patients were identified as the main sources of disagreement. Agreement on total judgment was comparable to results from similar studies in the literature using various algorithms, but in contrast to all these studies a higher percent of agreement (80%) was obtained with the use of the conventional criteria in this study.

Key words: Adverse drug reactions – Interrater agreement – Probability judgment – Algorithmic evaluation

Introduction

Since May 1979 adverse drug reactions (ADR) have been continuously assessed in inpatients of the psychiatric university hospitals of Berlin and Munich with AMÜP (**A**rznei**m**ittel-**ü**berwachung in der **P**sychiatrie), a drug surveillance program supported by the German Federal Health Agency [11, 25].

The probability of causal relationships between adverse events and drug therapy has been judged so far as "possible", "probable" or "definite" according to the criteria given by Seidl et al. [27] and Hurwitz and Wade [13]. However, several authors [9, 14, 17, 19, 21] demonstrated that comparisons of judgments based on these conventional assessment schemes resulted in considerable observer variation between different evaluators, even experts on ADR. Various algorithms were developed in consequence to standardize the decision-making process [5, 10, 18, 20, 23]. Applicability to clinical cases was demonstrated for all of them; when compared to implicit judgment, algorithmic procedures invariably resulted in better interrater agreement [2, 5, 14, 18, 21, 23]. Consequently, the use of algorithms in drug surveillance was recommended by these authors. Therefore, a study was designed to examine interobserver agreement on probability ratings between drug monitors from the two centers participating in the AMÜP program obtained by conventional "implicit" criteria and by an algorithm. The algorithm developed by Kramer et al. [20] was chosen for that study, as it seems to be the most differentiated of all algorithms proposed so far. To our knowledge no algorithm has ever been systematically used for assessment of ADR in psychiatric patients. So another aim of this study was to test the applicability of Kramer's algorithm in psychiatry.

Offprint requests to: R. Grohmann at the above address

Material and Methods

As reported earlier [11], all ADR leading to discontinuation of the imputed medication, so-called ADR grade III, have been systematically assessed in all inpatients of the two psychiatric hospitals participating in the drug surveillance program. For the purpose of this study 80 cases (30 from Berlin, 50 from Munich) were randomly selected from the total of 245 cases of ADR grade III assessed in 1982. Ratings were performed by raters L.G.S and R.G., who have cooperated as drug monitors within the drug surveillance program from its start in 1979. The original patients' records and ADR files, prepared in a way that ratings on probability and imputed medication were omitted, served as sources of information for the raters.

All cases were evaluated by both raters first according to the conventional criteria ("implicit judgment") used so far within the program and again 4 weeks later using Kramer's algorithm, blind to their prior judgment—as far as possible after this time interval.

With Kramer's algorithm a total score is derived by adding subscores on six different axes: I = previous general experience with the drug, II = alternative etiologic candidates, III = timing of events, IV = drug levels and evidence of overdose, V = dechallenge, VI = rechallenge. Total scores are related to categories of probability as follows:

total score < 0: no drug reaction
total score 0–3: possible drug reaction
total score 4–5: probable drug reaction
total score 6–7: definite drug reaction

Causal relationship between one event and the medication applied at the time of occurrence of the ADR had to be established. Judgment included a decision on the drug or drug combination to be held responsible for the ADR as well as a probability rating. All drugs given at the time of manifestation of the ADR and possibly responsible drug combinations were rated separately. With drug combinations a methodological problem arose in algorithmic evaluation; Kramer's procedural instructions do not provide sufficient information for cases of multiple drug therapy, in which not an interaction, but an additive effect of a combination of drugs with similar ADR profiles must be suspected to be responsible for an ADR. It did not appear justified to attribute the ADR merely to the drug scoring highest, as proposed by Kramer et al., in these cases. So it was decided to rate all drugs in combination which had obtained a total score of at least "zero" (the lowest possible score for a "possible" drug reaction) and to compare this combination's score to that of the highest scoring single drug.

Martindale's "The Extra Pharmacopoeia" [22] served as source of reference instead of the "American Physician's Desk Reference".

Data were analyzed concerning interrater agreement within the two methods and with respect to intrarater agreement between the two methods. Finally the ratings were compared to the judgment of the case conference in 1982 (before computer storage all ADR grade III assessed at the participating hospitals are finally evaluated at a case conference in Munich presided over by an experienced psychopharmacologist).

Statistical analysis of the data was performed by calculating percent of agreement and Cohen's [8] weighted kappa, a chance-adjusted measure which takes into account partial agreement. Differences of rates of agreement were tested by χ^2 tests applying McNemar's test for asymmetry (here: consistency) for dependent frequencies.

This paper will focus on the results of the algorithmic evaluation of ADR's with regard to applicability of Kramer's algorithm and to algorithmic interrater agreement.

Detailed results on the implicit evaluation and on the comparison of methods will be presented in forthcoming papers.

Results

In 63 cases (79% of all cases) more than one drug had been given at the time of the ADR resulting in an overall number of 230 individual drugs, among these 160 psychotropic and 70 nonpsychotropic drugs, and additional combinations that had to be evaluated. Most frequent drug groups were neuroleptics (62 exposures) and antidepressants (47 exposures); a neuroleptic and an antidepressant were used concomitantly in 18 cases, and two neuroleptics in 11 cases. As one example of combinations of drugs with similar ADR profiles combinations of at least two drugs with some anticholinergic activity (be it neuroleptic, antidepressant or anti-Parkinson drugs) were used in 20 cases (25% of cases).

Judgment on Imputed Drugs

In 69 of 80 cases (86%) partial agreement on the drugs to be incriminated was obtained, that is in 52 (83%) of the 63 multiple drug cases in addition to the 17 cases in which only one drug had been given at the time of the ADR. In all 11 cases with differences on imputed medication there was still some partial agreement, e.g. drug A versus combination of A + B or combination of A + B versus A + C.

A combination was imputed by both raters in 30% of all multiple drug cases with complete agreement on all elements of the combination in 15 cases and partial agreement in 4 cases. In another 11 cases only one rather imputed a drug combination.

Total Judgment

Probability ratings for the 69 cases with agreement on imputed medication are shown in Table 1. In 55 cases raters completely agreed in their final judgment, that is in 69% of all cases. Weighted kappa (0.618), measuring total agreement in drug-concordant cases, exceeded chance agreement considerably ($P < 0.001$).

Probability was judged by both raters as "possible" in 33%, as "probable" in 64% and as "definite" in only 3% of all

Table 1. Probability rating for the 69 cases out of 80 with concordance for imputed medication

		Rater L.G.S.			
		Possible	Probable	Definite	Total
Rater R.G.	Possible	18	9	0	27
	Probable	4	35	0	39
	Definite	0	1	2	3
	Total	22	45	2	69

Total agreement of judgment: 55 cases
Weighted kappa (kw) = 0.618, SE (kw) = 0.124, Z = 4.974
($P < 0.001$)

Table 2. Impact of number of drugs evaluated on interrater agreement (all 80 cases)

	No. of cases	Total agreement	
		No. of cases	(%)
1 drug evaluated	17	13	77%
2 drugs evaluated	24	19	79%
3 drugs evaluated	16	14	88%
≥ 4 drugs evaluated	23	9	39%
	80	55	67%

Table 3. Impact of type of ADR on interrater agreement (n = 80 cases)

	No. of cases	No. of cases with totally concordant judgment	
Toxic delirium	18	11	(61%)
Extrapyramidal disturbances	15	13	(87%)
Agitation	8	7	
Increased liver enzymes	6	1	
Exanthema	5	5	
Hypotonia	4	1	
Tremor (non-EPMS)	4	1	
Sedation	3	3	
Paranoid ideation	3	3	
Others	14	10	

Table 4. Concordant ratings per axis for all 249 medications evaluated by both raters (230 single drugs + 19 combinations)

	No. of concordant ratings	In percent of all 249 ratings	No of concordant ratings resulting in a subscore of					Concordant "zero" ratings ($n = 249$)
			−2	−1	0	+1	+2	
I = Previous general experience with drug	212	85%	n.a.	67	20	125		8%
II = Alternative etiologic candidates	187	75%	n.a.	17	128	3	39	51%
III = Timing of events	223	90%	0	n.a.	204	19	n.a.	82%
IV = Drug levels and evidence of overdose	248	99.6%	n.a.	0	248	0	n.a.	99.6%
V = Dechallenge	189	76%	n.a.	75	31	83	n.a.	12%
VI = Rechallenge	246	99%	n.a.	1	242	3		97%
Total score	124	50%						

Kw = 0.74 ($P < 0.001$), to test for concordance, McNemar $\chi^2 = 5.83$ ($P < 0.016$), to test for symmetry; n.a. = not applied

Table 5. Disagreement between raters per axis in all 249 medications evaluated by both raters

	No. of discordant ratings in which:		χ^2 McNemar	
	R.G. scored higher	L.G.S. scored higher		
Axis I	17	20	0.24	N.S.
Axis II	27	35	1.03	N.S.
Axis III	10	16	0.96	N.S.
Axis IV	0	1	0	
Axis V	23	37	3.27	$P < 0.07$
Axis VI	0	3	1.33	N.S.
Total score	49	76	5.83	$P < 0.016$

concordant cases. Differences in probability ratings never exceeded one step.

Percent agreement in total judgment did not differ significantly between single and multiple drug cases (76% versus 67%; $\chi^2 = 0.6$) nor between Berlin and Munich cases (73% versus 66%; $\chi^2 = 0.47$). Agreement rate in total judgment was comparable for cases with one, two, or three drugs given at the time of the ADR (76% to 88%) and dropped sharply only for cases with four or more drugs (39%) (Table 2). The number of cases per type of ADR was too small to allow statistically significant conclusions from percent agreement rate for different types of ADR, but some trends appeared noteworthy nevertheless (Table 3). For example agreement in *all* cases for skin rashes, sedation, delusions; still very good agreement for extrapyramidal reactions (i.e. Parkinsonism, acute dystonia, akathisia, and tardive dyskinesia), which are prominent ADR of neuroleptic drugs, and for psychic alterations summarized as "agitation"; low agreement for non-extrapyramidal motor system (EPMS) tremor, elevation of liver enzymes and for hypotonia.

Total Score and Subscore Ratings

In 38 cases (49% of all cases) raters agreed on total score for the imputed medication, in 34 cases (43% of all cases) they also agreed on all subscores.

A total of 230 individual drugs and 19 additional combinations were evaluated by both raters. They agreed on total score in 50% of these 249 drug ratings (Table 4). This agreement rate exceeded chance agreement by far (weighted kappa = 0.74, $P < 0.001$).

Evaluation of subscore ratings for all individual drugs and combinations (Table 4) revealed some difference in agreement rate between axes. Agreement was low for the axes II (alternative candidates) and V (dechallenge), whereas concordance rates of 90% or more were obtained for the other axes. But, as Table 4 illustrates, the high concordance rates resulted from a very high proportion of concordant "zero" scores on the axes III, IV, and VI (82% to 99.6%). Analysis of discordant ratings (Table 5) revealed that L.G.S. scored higher on total score than R.G. more frequently (McNemar $\chi^2 = 5.83$, $P < 0.016$). As to individual axes, only axis V (dechallenge) yielded a statistically significant difference in rating with L.G.S. scoring higher then R.G. more often again.

Analysis of the questions, which led to different subscore ratings most frequently (Table 6) identified as main sources of disagreement the question concerned with "maneuvers specifically directed against the adverse reaction" on axis V and the question whether the adverse event was "a change in a preexisting condition" on axis II.

On the average, raters needed 25 min per case for algorithmic evaluation.

Comparison to Implicit Judgment

Finally, Table 7 gives some basic results for agreement of algorithmic versus implicit judgment and versus judgment obtained by case conference. Using their conventional criteria raters agreed on total judgment in 80% of all cases. A similar percent agreement was obtained for both raters between their implicit judgment and that of the case conference, whereas algorithmic judgment of both raters was concordant with the case conference in only slightly more than 50%. Similarly, intrarater agreement between the two methods was only 53% and 58%. Weighted kappa (again calculated for medication concordant cases only for methodological reasons), however, signaled a percent of agreement exceeding chance agreement for all these comparisons.

Discussion

Agreement on total judgment in 69% of cases with an algorithmic evaluation, as obtained in this study, is comparable to the results of other studies using Kramer's algorithm, in which pairwise agreement ranged from 66% [15] to 73%–87% [14] and 63%–83% [21]. Only Naranjo et al. [23] reported a considerably higher percent of agreement of 83%–92% with

Table 6. Questions leading to different subscores most frequently by application of Kramer's algorithm

	Question no.	No. of discordant answers	Text of question
Axis I	1	13	Is the CM widely known and universally accepted as an adverse reaction to the suspected drug?
	5	25	Is the CM a change (exacerbation, recurrence, complication, or new manifestation) in a preexisting clinical condition, i.e., a condition present before the administration of the suspected drug?
Axis II	9	12	Is the CM consistent in quality and severity with any new alternative etiologic candidates other than a preexisting condition?
	11	20	Is the CM commonly seen with any of these alternative candidates?
Axis III	17	16	Given the type of CM; was the timing not only consistent with, but as expected for an adverse drug reaction to this drug?
Axis V	39/43	31	Was an agent or maneuver administered that was specifically directed against the CM and that usually produces the degree and rate of improvement observed in this case?

CM = Clinical manifestation—adverse event in this text

Table 7. Rates of agreement between raters and methods

		Partial agreement		Total agreement		Weighted Kappa
		No. of cases	%	No. of cases	%	
Interrater agreement:	With algorithm	69	86%	55	69%	0.618***
	With implicit judgment	74	93%	64	80%	0.409*
Agreement of rater versus case conference:	L.G.S. impl./c.c.	68	85%	63	79%	0.647**
	alg./c.c.	61	76%	44	55%	0.318*
	R.G. impl./c.c.	69	86%	64	80%	0.645**
	alg./c.c.	58	73%	41	51%	0.361**
Intrarater agreement:	L.G.S. impl./alg.	64	80%	46	58%	0.339*
	R.G. impl./alg.	63	79%	42	53%	0.310*

* = $P < 0.05$
** = $P < 0.01$
*** = $P < 0.001$

their algorithm, whereas Karch and Lasagna [18] found a comparable agreement of 71% using their own decision table. However, in all these studies the relation between one event and one given drug only was evaluated in contrast to this study. Only Blanc et al. [5] explicitly mentioned multiple drug therapy. Apparently they obtained agreement on imputed drugs and probability in about 40% of all cases only.

With Kramer's algorithm final judgment is the unweighted sum of judgments on six different axes, each of them dealing with one relevant aspect of ADR evaluation. This procedure allows closer analysis of disagreement on individual elements of the decision-making process.

On the *first axis*, concerned with previous general experience with the drug, raters agreed fairly well. The 15% disagreements mainly resulted from the question whether the ADR is "widely known" with a given drug. This question implies knowledge not only on the ADR itself but also on "general knowledge" on ADR's and is rather "jugdmental", as Hutchinson et al. [15] termed it.

Judgments on "alternative etiologic candidates" *(axis II)* were one important source of discordance here as in similar studies [14, 16, 21, 23]. Raters disagreed on axis II in 25% of all ratings. ADR symptoms often resemble characteristics of the underlying illness so that it may be difficult and requires a good deal of judgment to attribute such symptoms to drug therapy depending somewhat on the type of ADR. For instance, it would not be difficult to attribute Parkinsonism in a physically healthy young man to his neuroleptic treatment and not to his paranoid state. But if severe akinesia develops in a patient with catatonia, it may be very difficult to differentiate a deterioration of the illness to catatonic stupor from a neuroleptic syndrome. Similar difficulties arise, when a suspected toxic delirium has to be evaluated in a senile demented patient who may become confused during the night without any drug treatment, to cite only two examples. The different rates of agreement for different types of ADR found in this study provide evidence for this. The situation is even more intriguing, when several drugs with similar ADR profiles are given at the same time, which is common practice in the treatment of psychiatric inpatients [12, 26] and frequently occurred in the cases evaluated here, e.g., combinations of drugs with anticholinergic activity in 25% of all cases. In such cases each of the drugs as well as the combination of all suspect drugs represents a "good alternative candidate" to each other. The judgmental question, whether the adverse event is "common" with such an alternative candidate (axis II), was another source of disagreement mainly in such cases with different drugs representing alternative etiologic candidates. On *axis III* again the use of the judgmental term "as expected" in evaluation of time course mainly led to disagreement.

Analysis of the subscore values revealed that the very good concordance on *axis IV* resulted from almost 100% of "zero" scores on this axis. This failure of axis IV (drug levels and evidence of overdose) to yield any positive score is inevitable in view of the definition of an ADR used within this drug surveillance program as "any drug-related event that is undesired and unintended and occurs at a dosage appropriate for therapy or prophylaxis" according to the propositions of Seidl et al. [27] and Karch and Lasagna [16], thus explicitly excluding symptoms due to overdosage.

On *axis V* (dechallenge) the question which led to discordant answers most frequently was concerned with "maneuvers directed specifically against the adverse event". Such maneuvers are frequent in the course of psychotropic drug treatment in general and frequently occurred in the cases of this study. Anti-Parkinson drugs had been given for extrapyramidal symptoms, chlormethiazol for delirious states, tranquilizers for agitation. Such treatment courses may be accepted as rather specific maneuvers to treat the cited ADR in general. But there is no way of knowing exactly in an individual patient, which dosage and duration of such a treatment "usually produces the degree and rate of improvement observed in this case" (question 43 of Kramer).

Disagreement on that question was the most relevant individual factor contributing to overall discordance. Additional difficulties arose in evaluation of time course after discontinuation whenever such maneuvers against an adverse event are coupled with discontinuation of the suspect medication, which was frequent practice again.

Kramer's procedural instructions do not specify, whether discontinuation of a suspect drug, which may be regarded as a very specific action against an adverse reaction, is to be likewise considered in the evaluation of a drug which was not discontinued. So some additional disagreement resulted from different handling of that problem. In cases of alleged ADR persisting over a longer time period, which frequently happens especially in cases of Parkinsonism, akathisia, or increased liver enzymes, evaluation of dechallenge reaction, time course, and alternative candidates is complicated by a varying drug regimen with changing doses of different drugs, drugs added and withdrawn within that time period; the evaluator of ADR cases in psychiatric patients is often faced by this situation.

On axis VI again the good agreement resulted from nearly 100% of "zero" ratings, as rechallenge was not undertaken in most cases. The AMÜP drug surveillance program is conducted under conditions of routine drug treatment; readministration of a drug which has once appeared harmful, must be an exception in such circumstances.

In summary, and in view of an overall percent of agreement comparable to that obtained by other authors in medical and pediatric cases, Kramer's algorithm proved to be applicable for evaluation of ADR in psychiatric patients. In addition to the general problem of differentiation of an adverse event from illness-related changes, which will always require judgment by the rather, be it with or without application of an algorithm, two main sources for disagreement with the use of Kramer's algorithm in psychiatry were identified in this study.

First, the use of vague terms like "common", "as expected", "usually" led to disagreements, similarly to Hutchinson's results [15]. What relative frequency of occurrence in a given situation would made an event "common"? A better operationalisation of these terms obviously cannot solve the problem. Even if relative frequencies corresponding to "usually", "as expected" etc. were defined arbitrarily, one would have to rely for answers to the corresponding questions on exact knowledge about such frequencies of various events in various clinical situations. Such epidemiologic baseline data do not exist yet for psychiatric inpatients and to obtain that kind of information is one aim of the AMÜP program.

The second important source of disagreement with the use of Kramer's algorithm in psychistry is created by the frequent use of polypharmacy. Here an extension of procedural instructions should be feasible and would probably improve interrater agreement to some extent through identical application of the algorithm. In view of the multitude of problems created by the degree of polypharmacy, briefly outlined above, perfect agreement in multiple drug cases will obviously not be obtainable even with the best of algorithmic procedures.

The most striking result of this study was the higher percent agreement (80%) obtained with implicit judgment. This is in contrast to all other studies which had compared judgments based on conventional criteria and on algorithms [14, 17, 18, 21, 23].

Two main factors probably contributed to this apparently surprising result. First, using implicit judgment there is a remarkable tendency to attribute adverse events assessed as grade III within this drug surveillance program to the "probable" category, e.g., 80% of all ADR grade III assessed from 1979 to 1981 [25]. This was also reflected within this study where 60 of all 64 cases concordant with implicit judgment were rated as "probable". The different marginal distributions of rating categories observed by the two methods with far more "possible" ADR by algorithmic evaluation, are reflected in the increase of the weighted kappa value from implicit to algorithmic judgment.

A second and probably even more relevant explanation for the good implicit agreement between the two raters is the fact that both raters have been closely cooperating in the assessment of ADR for 5 years at the time of this study and, in course of that time, frequently discussed cases in which they had disagreed in the first place in order to come to a consensus. Considering this intensive training effect the remaining 20% disagreement may appear surprisingly high and one may speculate that after 5 years of continuous use of Kramer's algorithm the agreement rate would in fact exceed the one obtained implicitly.

But even if algorithms increase reproducibility of judgments on ADR this does not necessarily mean coming closer to "truth" about a drug-event relation. Several authors from the French "Centre de Pharmacovigilance" clearly demonstrated this. The "general knowledge on an ADR" relies on textbooks and prior reports on a given alleged ADR, but these sources are far from unequivocal. Reevaluation of 100 published ADR case reports with the algorithm of Dangoumau et al. [10] yielded judgments differing from those of the publishing authors by 47.5% [1]. Case reports mentioned in a textbook therefore must not be taken as sound evidence for an alleged ADR in any case. Furthermore, use of different textbooks leads to quite different results for "known ADRs'" as Begaud et al. [4] demonstrated in a comparative study including among others Martindale's "The Extra Pharmacopoiea", used in this study, and the "American Physicians' Desk Reference", recommended by Kramer et al. for use in their algorithm.

Judgments on ADR may change in the course of time, as Boisseau et al. [6] demonstrated. In his study a reevaluation of 82 ADR cases after 6 months resulted in a change of initial imputablility in 25% of all cases due to new evidence which had evolved in the meantime. The 20% disagreement of raters' implicit judgment with the case conference in this study, although both had been prominently involved in its decisions at that time, may be in part due to that factor.

Finally and most important the French group applied four different algorithms, those proposed by Karch and Lasagna [18], Kramer et al. [20], Blanc et al. [5], and Dangoumau et al. [10], to 100 cases and found unanimous verdicts in only 27% of all cases. Pairwise comparison of Kramer's with each of the other three algorithms yielded agreement in only about 50% of all cases [3]. Busto et al. [7] found a very high correlation of judgments based on their own algorithm to those obtained with Kramer's algorithm. However, in a second comparative study of five algorithms including the one proposed by Naranjo the French group obtained pairwise agreement of these two algorithms in 65% of cases only. Still, this was the best pairwise agreement of any two algorithms in this study [24]. The different importance attributed to single elements of the total judgment—like "previous experience with the drug", "time course" etc.—by the various algorithms was identified as the main source of disagreement between algorithms in the French studies. This may be illustrated by just one example. Whereas improvement of an ADR following the application of a specific agent against it leads to "0" instead of a possible "+1" rating on axis Vc in Kramer's algorithm, this fact is taken as a positive evidence for an ADR (+1) in the same way as improvement upon dechallenge (also +1) in Naranjo's algorithm. So the weight given to different criteria appears to be one crucial point of any algorithm.

The application of an algorithm is appropriate to pin-point reasons for disagreement between raters. It does not necessarily lead to an increase in interrater agreement in comparison to implicit judgment, if longstanding experience in their use favors the conventional criteria, as demonstrated in this study. Application of algorithms appears to be only one step towards real identification of an ADR. The next step, that is giving the proper weight to each of the various pieces of evidence, still remains to be taken, and it will require the joint efforts of all who work in the field of ADR assessment.

References

1. Albin H, Begaud B, Boisseau A, Dangoumau J (1980) Validation des publications d'effets indésirables par une methode d'imputabilité. Thérapie 35:571–576
2. Begaud B, Boisseau A, Albin H, Dangoumau J (1978) Imputabilité des effets indésirables des médicaments. Étude de 194 observations. Thérapie 33:383–389
3. Begaud B, Boisseau A, Dangoumau J (1981) Comparaison de quatre méthodes d'imputabilité des effets indésirables des médicaments. Thérapie 36:65–70
4. Begaud B, Pere JC, Dangoumau J (1981) Mise en œuvre d'un critére: la bibliographie. Thérapie 36:233–236
5. Blanc S, Leuenberger P, Berger JP, Brocke E, Schelling JL (1979) Judgments of trained observers on adverse drug reactions. Clin Pharmacol Ther 25:493–498
6. Boisseau A, Begaud B, Albin H, Dangoumau J (1980) Réévaluation du diagnostic d'effet indésirable des médicatments avec un recul de six mois. Thérapie 35:577–580
7. Busto U, Naranjo CA, Sellers EM (1981) Comparison of two recently published algorithms to assess the probability of adverse drug reactions. Clin Pharmacol Ther 32:236
8. Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
9. Dangoumau J, Begaud B, Boisseau A, Albin H (1980) Diagnostic des effets indésirables des médicaments (appréciations comparées de cliniciens et de pharmacologues cliniciens). Nouv Presse Med 9:1607–1609
10. Dangoumau J, Evreux JC, Jouglard J (1978) Méthode d'imputabilité des effets indésirables des médicaments. Thérapie 33:373–381
11. Grohmann R, Hippius H, Mueller-Oerlinghausen B, Ruether E, Scherer J, Schmidt LG, Strauss A, Wolf B (1984) Assessment of adverse drug reactions in psychiatric hospitals. Eur J Clin Pharmacol 26:727–734
12. Grohmann R, Strauss A, Gehr C, Ruether E, Hippius H (1980) Zur Praxis der klinischen Therapie mit Psychopharmaka. Pharmakopsychiatr 13:1–19
13. Hurwitz N, Wade OL (1969) Intensive hospital monitoring of adverse reactions to drugs. Br Med J 1:531–536
14. Hutchinson TA, Leventhal JM, Kramer MS, Karch FE, Lipman AG, Feinstein AR (1979) An algorithm for the operational assessment of adverse drug reactions. II. Demonstration of reproducibility and validity. JAMA 242:633–638
15. Hutchinson TA, Flegel KM, Ho Ping Kong H, Bloom WS, Kramer MS, Trummer EG (1983) Reasons for disagreement in the standardized assessment of suspected adverse drug reactions. Clin Pharmacol Ther 34:421–426
16. Karch FE, Lasagna L (1975) Adverse drug reactions—A critical review. JAMA 234:1236–1241
17. Karch FE, Smith CL, Kerzner B, Mazzullo JM, Weintraub M, Lasagna L (1976) Adverse drug reactions—a matter of opinion. Clin Pharmacol Ther 19:489–492
18. Karch FE, Lasagna L (1977) Toward the operational identification of adverse drug reactions. Clin Pharmacol Ther 21:247–254
19. Koch-Weser J, Sellers EM, Zacest R (1977) The ambiguity of adverse drug reactions. Eur J Clin Pharmacol 11:75–78
20. Kramer MS, Leventhal JM, Hutchinson TA, Feinstein AR (1979) An algorithm for the operational assessment of adverse drug reactions. I. Background, description and instructions for use. JAMA 242:623–632
21. Leventhal JM, Hutchinson TA, Kramer MS, Feinstein AR (1979) An algorithm for the operational assessment of adverse drug reactions. III. Results of tests among clinicians. JAMA 242:1991–1994
22. Martindale W (1982) The extra pharmacopoeia, 28th edn. The Pharmaceutical Press, London
23. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, Janecek E, Domecq C, Greenblatt DJ (1981) A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 30:239–244
24. Péré JC, Begaud B, Haramburu F, Albin H (1984) Méthodes d'étude des effets indésirables des médicaments. II. Profil et comparaison de cinq methodes d'imputabilité. Thérapie 39:369–378
25. Schmidt LG, Grohmann R, Helmchen H, Langscheid-Schmidt K, Mueller-Oerlinghausen B, Poser W, Ruether E, Scherer J, Strauss A, Wolf B (1984) Adverse drug reactions—An epidemiological study at psychiatric hospitals. Acta Psychiatr Scand 70:77–89
26. Schroeder NH, Caffey EM, Lorei TW (1977) Antipsychotic drug use: Physician prescribing practices in relation to current recommendations. Dis Nerv Syst 38:114–116
27. Seidl LG, Thornton GF, Smith JW, Cluff LE (1965) Epidemiologic studies of adverse drug reactions. Am J Public Health 55:1170–1175

Received June 28, 1985