Drug Information Journal, Vol. 20, pp. 357–361, 1986
Printed in the USA. Pergamon Journals Ltd.

0092-8615/86/$3.00 + .00

RELIABILITY OF ADVERSE DRUG REACTION ASSESSMENT IN PSYCHIATRIC INPATIENTS

R. GROHMANN,* MD, P. DIRSCHEDL,‡ MD, J. SCHERER,*
L. G. SCHMIDT,† MD, AND O. WUNDERLICH*

*Psychiatric Hospital of the University of Munich
†Psychiatric Hospital of the Free University of Berlin
‡Department of Medical Statistics and Data Processing of the University of Munich

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 inter-rater 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. 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 than with the use of the algorithm (69%). Possible explanations and consequences are discussed.

Key Words: Adverse drug reaction; Psychiatric inpatients; AMUP; Drug surveillance; Kramers algorithm

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.[6]

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*[17] and Hurwitz and Wade.[8] However, several authors[5,9,10,12,14] demonstrated that use of these conventional assessment schemes resulted in considerable observer variation between different raters and that algorithms developed in consequence to standardize the decision-making process increased rates of agreement in all their studies.[1,3,9,11,14,15] Therefore, it appeared necessary to examine interobserver agreement on probability ratings between drug monitors from the two centres participating in the AMÜP-program obtained by the conventional AMÜP-criteria and to compare these agreement rates to those obtainable with an algorithm.

The algorithm developed by KRAMER *et al*[13] was chosen for that study, as it seems

From the AMÜP-study supported by the Federal Health Agency (Bundesgesundheitsamt), Berlin, West Germany.

Correspondent address: Dr Renate Grohmann, Psychiatrische Universitätsklinik, Nußbaumstr. 7, D-8000 München 2, West Germany.

to be the most differentiated of all algorithms proposed so far.

MATERIAL AND METHODS

Eighty cases (30 from Berlin, 50 from Munich), in which an ADR had led to discontinuation of a suspect medication were randomly selected from the total of 245 such cases assessed in 1982.

All cases were evaluated by two raters (L.G.S. and R.G., who have cooperated as drug monitors within the drug surveillance program from its start in 1979), first with the conventional AMÜP-criteria ("implicit judgment") and again 4 weeks later using KRAMER'S algorithm, blind to their prior judgment – as far as possible after this time interval.

Judgment included a decision on probability rating and on the drug or drug combination to be held responsible for the ADR. All drugs given at the time of manifestation of the ADR were rated separately as well as possibly responsible drug combinations.

Inter-rater agreement within the two methods and intra-rater agreement between the two methods were analyzed. Finally the decisions of the case conference in 1982 (before computer storage all ADR assessed in the participating hospitals are finally evaluated at a case conference in Munich presided over by an experienced psychopharmacologist) were compared to the implicit and algorithmic judgments. Statistical analysis of the data included calculating percent of agreement and COHEN'S weighted kappa,[4] a chance-adjusted measure which takes into account partial agreement.

RESULTS

Table 1 shows the overall agreement rates. 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. In contrast interrater agreement with the algorithm was only 69% and algorithmic judgments of both raters were concordant with the case conference in only slightly

TABLE 1
Rates of Agreement between Raters and Methods

	Partial agreement no. of cases	Percent (%)	Total agreement no. of cases	Percent (%)	Weighted Kappa
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*

* = $p < 0.05$.
† = $p < 0.01$.
‡ = $p < 0.001$.

more than 50%. Similarly, intrarater-agreement between the methods was only 53% and 58%. Weighted kappa (calculated for medication-concordant cases only for methodological reasons), however, signals a percent of agreement exceeding chance agreement for all these comparisons.

Analyzed according to number of drugs applied at the time of the ADR agreement rates decrease with both methods, when four or more drugs are concerned, but this decrease is far more distinct for the algorithmic evaluation (Table 2). The high rate of discordance with the algorithm in cases of multiple drug therapy was due not only to the scant procedural instructions provided by the authors of the algorithm for such cases leading to somewhat different application of the algorithm, but also to the logical problems resulting from combinations of drugs with similar side effects when each drug has to be examined as an alternative to each other on axis II of the algorithm exploring the alternative etiologic candidates.

Comparison of the degrees of probability attributed to the cases with concordant judgment on incriminated drugs reveals that the implicit method led to a high number of "probable" ratings, whereas the algorithm resulted in far more judgments of a "possible" relationship (Table 3). The resulting difference in marginal distribution was responsible for the higher kappa-values for the algorithmic inter-rater comparison (see Table 1) in spite of a lower percent of agreement relative to the implicit method. Kramer's algorithm allows closer analysis of the factors leading to a final probability rating. It consists of six different axes, each of them dealing with one element of the causality assessment. Subscores obtained on these axes are added up to a total score which is transformed into a degree of probability according to a prefixed scheme.

Table 4 gives the concordant subscores per axis in percent of all ratings for the 249 drugs and drug combinations evaluated by both raters. Agreement rates of 90% and more were obtained for axis III, IV, and VI, but, as Table 4 also demonstrates, this very good agreement resulted from concordant zero-ratings mostly. This can be explained by the nature of these axes: On axis III a time course "as expected" is required for a positive score; with ADR in psychiatry this cannot be stated positively as a rule. Axis IV asks for evidence of overdosage which is explicitly excluded in the AMÜP-study's definition of an ADR as in most other drug surveillance programs, so that necessarily no positive score can be obtained on this axis. Axis VI deals with rechallenge, which is only rarely undertaken in a study operating under conditions of everyday treatment in which avoidance of once harmful drugs is aimed at. This necessarily reduces the chance to achieve a total score required for a "probable" rating (a total score of +4 or +5).

TABLE 2
Agreement Rate and Number of Drugs (All 80 Cases)

	No. of cases	Agreement rate impl.	Agreement rate alg.
1 drug	17	88%	77%
2 drugs	24	88%	79%
3 drugs	16	88%	88%
≥4 drugs	23	78%	39%

Case-wise comparison of the raters' judgments with the two methods revealed that in 16 cases raters agreed within each method but concordantly disagreed between the two methods. In all these cases the relationship was rated as only "possible" with the algorithm and as "probable" with the implicit method by both raters. The fact that a preexisting illness predisposing to an ADR – like organic brain damage to toxic delirium or liver damage to drug-induced increase of transaminases – reduces the maximum score to be obtained on axis II of the algorithm could be identified as the source of this systematic discrepancy between the methods as the implicit method does not take such preexisting risk factors as negative evidence in ADR evaluation.

TABLE 3
Probability Rating with Implicit Judgment (74 Drug-Concordant Cases) and with the Algorithm (69 Drug-Concordant Cases)

		Implicit Rater L.G.S. Poss.	Prob.	Def.	Total	Algorithm Rater L.G.S. Poss.	Prob.	Def.	Total
Rater R.G.	Poss.	0	4	0	4	18	9	0	27
	Prob.	5	60	0	65	4	35	0	39
	Def.	0	1	4	5	0	1	2	3
	Total	5	65	4	74	22	45	2	69

CONCLUSIONS

In contrast to earlier results of other authors[1,3,9,11,14,15] this study did not demonstrate a superiority of algorithmic evaluation with regard to the degree of interrater agreement, as the percent of agreement obtained by the implicit method exceeded that of the algorithm. This is probably due in part to the long-standing experience of the raters with the implicit method. But data analysis identified specific difficulties of the algorithmic procedure in cases of multiple drug therapy as one source of disagreement which weighs heavily in psychiatry where polypharmacy is common practice.[7,16] It should be mentioned here that with one exception[3] inter-rater reliability studies so far evaluated the relationship between an adverse event and one given drug only.

In addition, the results of this study clearly demonstrate that the final judgments depend not only on the known facts but also on the way in which the decision-making process is constructed as the weight given to single facts varies. This could be seen here from the example of the factor "preexisting illness". Similarly Begaud *et al*[2] found unanimous verdicts in only 27% of 100 cases evaluated by four different algorithms including that of Kramer *et al*; pairwise comparisons of Kramer's with each of the other three algorithms revealed agreement in only about 50% of all cases. This discordance between different algorithms also had to be attributed to different weight given to single elements of the decision-making process. Thus the main problem in causality assessment of ADR still awaiting a solution apparently is to decide

TABLE 4.
Concordant Ratings Per Axis

	Concordant ratings in percent	Concordant ratings score = 0
I = gen. experience	85%	8%
II = alternat.etiol.cand.	75%	51%
III = time course	90%	82%
IV = evidence overdose	99.6%	99.6%
V = dechallenge	76%	12%
VI = rechallenge	99%	97%

which is the proper weight to be attributed to the various pieces of evidence.

REFERENCES

1. Begaud B, Boisseau A, Albin H, Dangoumau J: Imputabilité des effects indésirables des médicaments. Etude de 194 observations. *Thérapie* 1978; 33:383-389.
2. Begaud B, Boisseau A, Dangoumau J: Comparaison de quatre méthodes d'imputabilité effets indésirables des médicaments. *Thérapie* 1981; 36: 65-70.
3. Blanc S, Leuenberger P, Berger JP, Brocke E, Schelling JL: Judgments of trained observers on adverse drug reactions. *Clin Pharm Ther* 1979; 25:493-498.
4. Cohen J: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70:213-220.
5. Dangoumau J, Begaud B, Boisseau A, Albin H: Diagnostic des effets indésirables des médicaments (appréciations comparées de cliniciens et de pharmacologues cliniciens). *Nouv Presse Med* 1980; 9:1607-1609.
6. Grohmann R, Hippius H, Mueller-Oerlinghausen B, Ruether E, Scherer J, Schmidt LG, Strauss A, Wolf B: Assessment of Adverse Drug Reactions in Psychiatric Hospitals. *Eur J Clin Pharmacol* 1984; 26:727-734.
7. Grohmann R, Strauss A, Gehr C, Ruether E, Hippius H: Zur Praxis der klinischen Therapie mit Psychopharmaka. *Pharmakopsychiat* 1980; 13:1-19.
8. Hurwitz N, Wade OL: Intensive Hospital Monitoring of Adverse Reactions to Drugs. *Brit Med J* 1969; 1:531-536.
9. Hutchinson TA, Leventhal JM, Kramer MS, Karch FE, Lipman AG, Feinstein AR: An Algorithm for the Operational Assessment of Adverse Drug Reactions II. Demonstration of Reproducibility and Validity. *JAMA* 1979; 242:633-638.
10. Karch FE, Smith CL, Kerzner B, Mazzullo JM, Weintraub M, Lasagna L: Adverse drug reactions —a matter of opinion. *Clin Pharm Ther* 1976; 19:489-492.
11. Karch FE, Lasagna L: Toward the operational identification of adverse drug reactions. *Clin Pharm Ther* 1977; 21:247-254.
12. Koch-Weser J, Sellers EM, Zacest R: The Ambiguity of Adverse Drug Reactions. *Eur J Clin Pharmacol* 1977; 11:75-78.
13. Kramer MS, Leventhal JM, Hutchinson TA, Feinstein AR: An Algorithm for the Operational Assessment of Adverse Drug Reactions. I. Background, Description and Instructions for Use. *JAMA* 1979; 242:623-632.
14. Leventhal JM, Hutchinson TA, Kramer MS, Feinstein AR: An Algorithm for the Operational Assessment of Adverse Drug Reactions III. Results of Tests Among Clinicians. *JAMA* 1979; 242: 1991-1994.
15. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, Janecek E, Domecq C, Greenblatt DJ: A method for estimating the probability of adverse drug reactions. *Clin Pharm Ther* 1981; 30:239-244.
16. Schroeder NH, Caffey EM, Lorei TW: Antipsychotic Drug Use: Physician Prescribing Practices in Relation to Current Recommendations. *Dis Nerv Syst* 1977; 38:114-116.
17. Seidl LG, Thornton GF, Smith JW, Cluff LE: Epidemiologic studies of adverse drug reactions. *Am J Public Health* 1965; 55:1170-1175.