A method for estimating the probability of adverse drug reactions

The estimation of the probability that a drug caused an adverse clinical event is usually based on clinical judgment. Lack of a method for establishing causality generates large between-raters and within-raters variability in assessment. Using the conventional categories and definitions of definite, probable, possible, and doubtful adverse drug reactions (ADRs), the between-raters agreement of two physicians and four pharmacists who independently assessed 63 randomly selected alleged ADRs was 38% to 63%, kappa (κ, a chance-corrected index of agreement) varied from 0.21 to 0.40, and the intraclass correlation coefficient of reliability (R[est]) was 0.49. Six (testing) and 22 wk (retesting) later the same observers independently reanalyzed the 63 cases by assigning a weighted score (ADR probability scale) to each of the components that must be considered in establishing causal associations between drug(s) and adverse events (e.g., temporal sequence). The cases were randomized to minimize the influence of learning. The event was assigned a probability category from the total score. The between-raters reliability (range: percent agreement = 83% to 92%; κ = 0.69 to 0.86; r = 0.91 to 0.95; R(est) = 0.92) and within-raters reliability (range: percent agreement = 80% to 97%; κ = 0.64 to 0.95; r = 0.91 to 0.98) improved ($p < 0.001$). The between-raters reliability was maintained on retesting (range: r = 0.84 to 0.94; R(est) = 0.87). The between-raters reliability of three attending physicians who independently assessed 28 other prospectively collected cases of alleged ADRs was very high (range: r = 0.76 to 0.87; R(est) = 0.80). It was also shown that the ADR probability scale has consensual, content, and concurrent validity. This systematic method offers a sensitive way to monitor ADRs and may be applicable to postmarketing drug surveillance.

C. A. Naranjo, M.D., U. Busto, Pharm.D., E. M. Sellers, M.D., Ph.D., P. Sandor, M.D., I. Ruiz, Pharm.D.,* E. A. Roberts, M.D., E. Janecek, B.Sc. Phm., C. Domecq, Pharm.D.,* and D. J. Greenblatt, M.D.** *Toronto, Ontario*
Clinical Pharmacology Program, Addiction Research Foundation Clinical Institute, and Departments of Medicine and Pharmacology, University of Toronto

Received for publication April 15, 1980.

Accepted for publication March 20, 1981.

Reprint requests to: C. A. Naranjo, M.D., Clinical Pharmacology Program, Addiction Research Foundation Clinical Institute, 33 Russell St., Toronto, Ontario M5S 2S1, Canada.

*Clinical Pharmacy Group, Faculty of Chemical Sciences, Universidad de Chile, Santiago, Chile.

**Division of Clinical Pharmacology, New England Medical Centre Hospital, Boston, MA.

The most important problem in assessing adverse drug reactions (ADRs) is whether there is a causal relationship between the drug and the untoward clinical event. The use of the conventional definitions and probabilities of definite, probable, possible, and doubtful ADRs[5] generates wide variability in assessment. Koch-

0009-9236/81/080239+07$00.70/0

Table I. *ADR probability scale*

To assess the adverse drug reaction, please answer the following questionnaire and give the pertinent score.

	Yes	*No*	*Do not know*	*Score*
1. Are there previous *conclusive* reports on this reaction?	+1	0	0	
2. Did the adverse event appear after the suspected drug was administered?	+2	−1	0	
3. Did the adverse reaction improve when the drug was discontinued or a *specific* antagonist was administered?	+1	0	0	
4. Did the adverse reaction reappear when the drug was readministered?	+2	−1	0	
5. Are there alternative causes (other than the drug) that could on their own have caused the reaction?	−1	+2	0	
6. Did the reaction reappear when a placebo was given?	−1	+1	0	
7. Was the drug detected in the blood (or other fluids) in concentrations known to be toxic?	+1	0	0	
8. Was the reaction more severe when the dose was increased, or less severe when the dose was decreased?	+1	0	0	
9. Did the patient have a similar reaction to the same or similar drugs in *any* previous exposure?	+1	0	0	
10. Was the adverse event confirmed by any objective evidence?	+1	0	0	
			Total score	

Weser et al.[8] found that clinical pharmacologists frequently disagreed when analyzing the causality of ADRs, and others[1, 7] have come to similar conclusions. Manifestations of ADRs are nonspecific. The suspected drug is usually confounded with other causes, and often the adverse clinical event cannot be distinguished from manifestations of the disease. Recently there have been attempts to systematize the assessment of causality of ADRs, applying operational definitions such as those proposed by Karch and Lasagna[6] and by Kramer et al.[9] The application of these methods in routine clinical practice has been limited, perhaps because they are too detailed and time consuming. We developed a simple method to assess the causality of ADRs in a variety of clinical situations, and its systematic application to different cases of alleged ADRs has provided reliable answers.

Materials and methods

To test the reliability and validity of the ADR probability scale (Table I) several studies were conducted. In the main study, on three occasions (phases 1, 2, and 3) six observers (two physicians and four pharmacists) independently assessed 63 randomly selected alleged ADRs. These cases composed a stratified random sample (18.8%) of 335 cases of ADRs published during 1978 in the *British Medical Journal* (22 cases), *Lancet* (17 cases), *Annals of Internal Medicine* (12 cases), *Journal of the American Medical Association* (8 cases), and *New England Journal of Medicine* (4 cases).* The cases were randomized to minimize learning, and the sequence was kept blind to the observers.

In the first assessment (phase 1) an "adverse drug reaction" (ADR) was defined as any noxious, unintended, and undesired effect of a drug after doses used in humans for prophylaxis, diagnosis, or therapy. This definition excludes therapeutic failures, intentional and accidental poisoning, and drug abuse.[16] The probability that the adverse event was related to drug therapy was classified as definite, probable, possible, or doubtful.[5, 12] A "definite" reaction was one that (1) followed a reasonable temporal sequence after a drug or in which a toxic drug level had been established in body fluids or tis-

*A list of the reports will be provided on request.

sues, (2) followed a recognized response to the suspected drug, and (3) was confirmed by improvement on withdrawing the drug and reappeared on reexposure. A "probable" reaction (1) followed a reasonable temporal sequence after a drug, (2) followed a recognized response to the suspected drug, (3) was confirmed by withdrawal but not by exposure to the drug, and (4) could not be reasonably explained by the known characteristics of the patient's clinical state. A "possible" reaction (1) followed a temporal sequence after a drug, (2) possibly followed a recognized pattern to the suspected drug, and (3) could be explained by characteristics of the patient's disease. A reaction was defined as "doubtful" if it was likely related to factors other than a drug.

Six weeks later the 63 cases were reordered randomly and reanalyzed (phase 2). The observers independently assigned a weighted score to the components used to establish a causal association between drugs and adverse events (temporal sequence, pattern of response, withdrawal, reexposure, alternative causes, placebo response, drug levels in body fluids or tissues, dose-response relationship, previous patient experience with the drug, and confirmation by objective evidence). These factors were analyzed and scored using the ADR probability scale (Table I). Each question could be answered positive (yes), negative (no), or unknown or inapplicable (do not know). The raters were instructed to use the questionnaire for about 20 min.* The ADR was assigned to a probability category from the total score as follows: definite ≥ 9, probable 5 to 8, possible 1 to 4, doubtful ≤ 0. The between-raters reliability to use the categorical classification of ADR probability was measured using percent agreement and kappa (κ, a chance-corrected index of agreement).[14] Kappa was calculated as follows:

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

where P_o = proportion of observed agreement and P_c = proportion of agreement expected by chance. Kappa ranged from -1 (complete disagreement) to $+1$ (perfect agreement). Correlation coefficients between ADR scores were also used to test between-raters and within-raters reliability in phases 2 and 3. The intraclass correlation coefficient of reliability (R[est]) was also calculated:

$$R(est) = \frac{S_s^2}{S_s^2 + S_r^2 + S_e^2}$$

where S_s^2 = variance from the cases, S_r^2 = variance generated by the raters, and S_e^2 = residual variance or error. This coefficient is the ratio of the variance associated with true case-to-case variability to the sum of all the components of variance. R(est) varies from zero (i.e., no intercase variation is detected by the ratings, the ratings are the result only of measurement error and between-rater differences) to a maximum of unity (i.e., intercase variation is correctly detected by the ratings, there is no contamination by measurement error or rater-to-rater variation).[14] The R(est) was calculated in phase 1, assuming a score of 1 (doubtful), 2 (possible), 3 (probable), or 4 (definite). The actual ADR scores were used in phases 2 and 3.

To determine whether the improvement in reliability found in phase 2 had occurred by chance the cases were again reordered randomly and reanalyzed independently by the six raters 4 mo later (phase 3). This allowed us to assess within-rater and between-rater retest reliability. The between-rater reliability of practicing physicians was also tested. Three attending physicians independently rated 28 other prospectively collected cases of alleged ADR observed in the Toronto Western Hospital.

Validity. To establish validity comparison with a standard is necessary. Because there is no method that can determine which adverse events are truly ADR, we studied the validity of the ADR probability scale in several ways. Consensual validity was tested as follows. (1) The consensus assessment of three "experts" (C. A. N., E. M. S., D. J. G.) using the conventional categories of ADR probabilities was the external standard with which physicians-pharmacists assessments were compared. Their expertise is supported by publications.[8, 10, 12]

*An appendix with instructions for using our ADR probability scale will be supplied with reprints and will also be available from the National Auxiliary Publication Service, American Society of Information Services, 1010 16th St. N.W., Washington, D.C. 20036.

Table II. *Between-raters agreement*

Pairs of raters	Phase 1		Phase 2			Phase 3
	%	κ	%	κ	r	r
R01-R02	52	0.35	83	0.69	0.93	0.85
R04	56	0.37	83	0.70	0.93	0.84
R06	44	0.22	86	0.75	0.92	0.94
R08	49	0.32	87	0.77	0.94	0.89
R10	54	0.35	84	0.72	0.93	0.91
R02-R04	54	0.31	83	0.71	0.91	0.87
R06	49	0.29	89	0.80	0.94	0.87
R08	48	0.32	86	0.75	0.93	0.89
R10	52	0.29	90	0.83	0.95	0.90
R04-R06	54	0.35	84	0.72	0.91	0.87
R08	48	0.36	83	0.70	0.93	0.87
R10	57	0.36	83	0.71	0.91	0.86
R06-R08	46	0.27	92	0.86	0.94	0.91
R10	54	0.35	90	0.83	0.92	0.93
R08-R10	41	0.21	86	0.77	0.94	0.93
Intraclass correlation coefficient of reliability	R(est) = 0.49		R(est) = 0.92			R(est) = 0.87

Table III. *Within-raters agreement*

Rater	Phase 1 vs. phase 2		Phase 1 vs. phase 3		Phase 2 vs. phase 3		
	%	κ	%	κ	%	κ	r
R01	43	0.23	38	0.16	92	0.85	0.96
R02	67	0.47	63	0.50	86	0.75	0.91
R04	54	0.28	48	0.19	80	0.64	0.94
R06	44	0.22	44	0.22	97	0.95	0.98
R08	36	0.17	43	0.25	87	0.78	0.93
R10	51	0.26	57	0.38	87	0.79	0.97

(2) One of the experts (C. A. N.) assessed the reactions using the ADR probability scale, and his ratings were compared with those by the physicians-pharmacists in phase 2. Content validity was tested in the 63 reported cases and in the 28 prospectively collected cases, comparing the variations in the ADR scores of reactions considered possible, probable, or definite and those classified as definite nondrug adverse events. Concurrent validity was tested by comparing the correlation of the scores of the 63 ADRs obtained by our method with those derived by the algorithm described by Kramer et al.[2, 9]

Results

Table II shows that there was poor between-raters agreement when the conventional definitions of ADRs were used (phase 1). Percent agreement ranged from 41% to 57% (kappa = 0.21 to 0.37, R(est) = 0.49). When the observers applied the ADR probability scale (phase 2) there was a rise in percent agreement (83% to 92%), κ (0.69 to 0.86), and r (0.91 to 0.95) (sign test, $p < 0.001$). The intraclass correlation coefficient of reliability (R[est] = 0.92) indicates high reproducibility. The between-raters reliability was maintained on phase 3 retesting (r = 0.84 to 0.93, R(est) = 0.87). The high within-raters reliability using the ADR probability scale (phase 2 versus phase 3) is shown in Table III. The percent agreement ranged from 80% to 97% (κ = 0.64 to 0.95, r = 0.91 to 0.98). The between-raters reliability of the three attending physicians who rated the 28 prospectively collected ADRs was also high (r = 0.76 to 0.87, R[est] = 0.80).

Validity. Percent agreement between the consensus of experts and the physicians-pharmacists assessments ranged from 79% to 84% (κ = 0.64 to 0.71). Percent agreement with the expert (C. A. N.) who used the ADR probability scale ranged from 86% to 95% (κ = 0.75 to

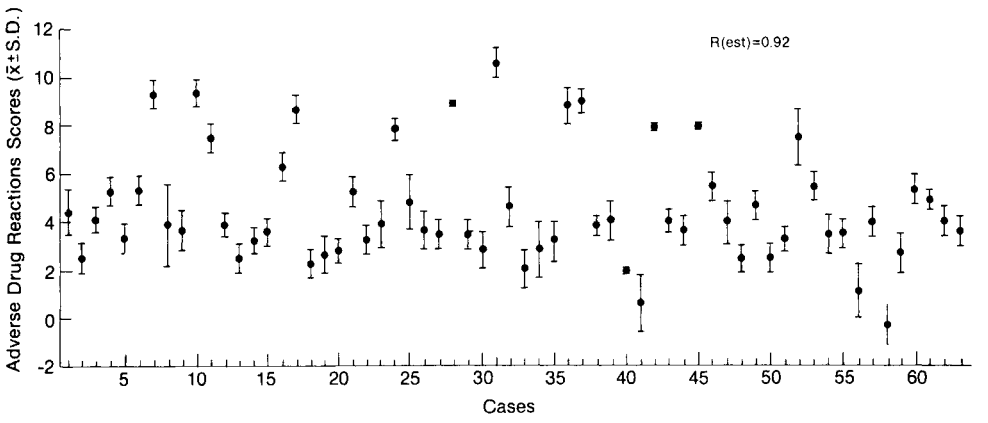

Fig. 1. Distribution of ADR scores in 63 cases of alleged ADRs.

0.91, r = 0.94 to 0.96). The ADR scores obtained rating the 63 reported cases with our method correlated with those derived using the algorithm described by Kramer et al.[2] (r = 0.82, $p < 0.001$).

Discussion

Our data indicated a marked improvement in between-raters and within-raters agreement when the adverse events were assessed with our ADR probability scale. The intraclass correlation coefficient of reliability (R[est] = 0.92) suggests that the method can discriminate ADRs of different probabilities. The reproducibility was maintained on retesting, and results of the same order were obtained when physicians rated a different set of prospectively collected cases of ADR. The ADR probability scale is a simple questionnaire that can be answered rapidly.

A major problem in drug-monitoring studies is lack of a reliable method of assessing the causal relation between drugs and adverse events. Such a method is needed because the incidence of adverse events can be estimated only from cases identified as definite or probable ADRs.[5] Our data and those of others have demonstrated large interobserver variations in assessments when the conventional categorical definitions of probability of ADRs were used.[1, 7, 8] Our ADR probability scale led to improved reproducibility in assessments. Using our scale, pairs of raters had scores that were within the same diagnostic category or only one category apart. When there was disagreement it was usually not substantial, as indicated by the small standard deviation of the ADR scores (Fig. 1) and the high correlation coefficients between scores (Tables II and III). A 3-point between-raters disagreement occurred in only one very complicated case.

It is possible that high reproducibility could occur without using the ADR probability scale, but the poor within-raters (phase 1/2 and phase 1/3; Table III) and between-raters agreements (phase 1; Table II) using the conventional definitions rule out this possibility. Perhaps the high agreement occurred because the 63 cases were selected from published reports and included only three ADR categories (possible, probable, and definite), which generated spuriously high reproducibility, but this seems unlikely. Fig. 1 shows that the cases represented a broad spectrum of ADRs (scores ranged from −2 to +12). The good correlation between the actual ADR scores reflecting between-raters reliability (r = 0.91 to 0.95; Table II) and within-raters reliability (r = 0.91 to 0.98; Table III) and the high κ values suggest that the ADR probability scale was the basis of a genuine improvement in reproducibility. When attending physicians used our method to rate a different set of reactions the between-raters agreement was good (R[est] = 0.80).

Using the ADR probability scale we were also able to identify the origin of the interrater

disagreements. The assessment of question 5 (alternative causes) led to the most disagreement. In view of the complex clinical situations and the differences in training of the observers, this should have been anticipated. Pharmacists in general were more likely to answer "I do not know" to this question. Hutchinson et al.[4] found that this could be a major source of disagreement even though very detailed instructions were given. In some complicated cases no algorithm can substitute for clinical experience.

Even though the reproducibility of an instrument is important, its validity must also be considered. The observers could agree among themselves, but they could also all be wrong. In cases of adverse events there is no definite standard against which to test the validity of new operational definitions of ADRs. We therefore assessed the validity of our ADR probability scale in several ways. The agreement of the six raters with the consensus of three experts was very high, suggesting that our instrument has consensual validity. Although the experts may not always accurately classify reactions, the probability that the consensus of three experts would be completely wrong all the time is small. The high agreement between the physicians-pharmacists and one of the experts using the ADR probability scale also indicates consensual validity. The concurrent validity of our instrument is suggested by the good correlation between the ADR scores generated by our method and those of another recently published algorithm.[2] The negative scores in the definite nondrug adverse events and the positive scores in the "true" ADR indicate that our method had content validity. Our findings indicate that our ADR probability scale is reliable and valid.

Important potential applications of the ADR probability scale are the analysis of adverse drug-related events published in medical journals as well as the assessment of reports submitted to national drug monitoring centers. Many countries are interested in developing postmarketing drug surveillance programs.[3] The reliability of the ADR assessments in case studies could improve if operational definitions such as ours and similar procedures are used.[15] Advantages of our method are simplicity and wide applicability. Some minor modifications may be required in special circumstances. In analyzing adverse drug interactions suspect interacting drugs rather than a particular drug must be assessed. When a patient receives several drugs at the same time the ADR scale must be applied to each of the possible causes; the most likely will be the drug with the highest score. In reactions that appear during drug withdrawal, withdrawal corresponds to reinstituting treatment and repetition corresponds to discontinuing the suspect drug.

The conventional classification of definite, probable, possible, and doubtful ADRs, as proposed by Seidl et al.[13] in 1966, assumes four discrete categories for which there is no empirical demonstration. It is therefore reasonable to postulate that some of the unreliability of the conventional definitions or operational definitions of ADRs could result, because such categories are not unique (i.e., the unreliability could reflect the overlap between nondiscrete categories). Thus the higher correlation coefficients of the actual ADR scores ($r = 0.91$ to 0.94), as compared with the kappa values when using four categories ($\kappa = 0.69$ to 0.83) (Table II), support this view and indicate the need to characterize the probability spectrum of ADRs empirically. We suggest that it is preferable to classify the probability using the actual ADR scores by our and similar operational methods.[11]

Notwithstanding our encouraging results, it is unrealistic to expect that our relatively simple procedure will solve all the complex problems of identification and classification of ADRs. Further experience will provide the rationale for refinements and improvements and will confirm its utility in clinical practice. Our findings suggest that its systematic application can improve the quality of the assessment of ADRs in a variety of clinical situations.

The collaboration of Doctors M. Spino, H. Wang, and M. Rudyk and of S. Schachter, B.Sc., in some of the phases of the study is gratefully acknowledged.

References

1. Blanc S, Leuenberger P, Berger JP, Brooke EM, Schelling JL: Judgments of trained observers on adverse drug reactions. CLIN PHARMACOL THER **25**:493-498, 1979.

2. Busto U, Naranjo CA, Sellers EM: Comparison of two recently published algorithms to assess the probability of adverse drug reactions. CLIN PHARMACOL THER **29:**236, 1981.
3. Gross FH, Inman WHW: Drug monitoring. New York, 1977, Academic Press.
4. 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 **242:**633-638, 1979.
5. Karch FE, Lasagna L: Adverse drug reactions: A critical review. JAMA **234:**1236-1241, 1975.
6. Karch FE, Lasagna L: Toward the operational identification of adverse drug reactions. CLIN PHARMACOL THER **21:**247-254, 1977.
7. Karch FE, Smith CL, Kerzner B, Mazzullo JM, Weintraub M, Lasagna L: Adverse drug reactions: A matter of opinion. CLIN PHARMACOL THER **19:**489-492, 1976.
8. Koch-Weser J, Sellers EM, Zacest R: The ambiguity of adverse drug reactions. Eur J Clin Pharmacol **11:**75-78, 1977.
9. 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 **242:**623-632, 1979.
10. Miller RR, Greenblatt DJ: Drug effects in hospitalized patients. New York, 1976, John Wiley & Sons.
11. Naranjo CA, Busto U, Abel JG, Sellers EM: Empiric delineation of the probability spectrum of adverse drug reactions. CLIN PHARMACOL THER **29:**267-268, 1981.
12. Naranjo CA, Pontigo E, Valdenegro C, Gonzalez G, Ruiz I, Busto U: Furosemide-induced adverse reactions in cirrhosis of the liver. CLIN PHARMACOL THER **25:**154-160, 1979.
13. Seidl LG, Thornton GF, Smith JW, Cluff LE: Studies on the epidemiology of adverse drug reactions. III. Reactions in patients on a general medical service. Johns Hopkins Med J **119:**299-315, 1966.
14. Spitzer RL, Fleiss JL, Endicott J: Problems of classification: Reliability and validity, *in* Lipton MA, DiMascio A, Killam DF, editors: Psychopharmacology: A generation of progress. New York, 1978, Raven Press, pp. 857-869.
15. Wardell WM, Tsianco MC, Anavekar SN, Davis HT: Postmarketing surveillance of new drugs. II. Case studies. J Clin Pharmacol **19:**169-184, 1979.
16. World Health Organization: International drug monitoring: The role of the hospital. WHO Tech Rep Ser No. 425, 1969