Br. J. clin. Pharmac. (1982), **13**, 223–227

COMPARISON OF TWO RECENTLY PUBLISHED ALGORITHMS FOR ASSESSING THE PROBABILITY OF ADVERSE DRUG REACTIONS

U. BUSTO, C.A. NARANJO & E.M. SELLERS

Clinical Pharmacology Program, Addiction Research Foundation,
Clinical Institute, 33 Russell Street, Toronto, Ontario M5S 2S1, Canada

1 A simple valid and reliable method for estimating the probability of adverse drug reactions (adverse drug reactions probability scale, APS) has been recently described (Naranjo *et al.*, 1981a).

2 The results using APS were compared to those obtained with another more detailed algorithm (adverse reactions scoring system, ASS) described by Kramer *et al.* (1979).

3 Sixty-three randomly selected adverse drug reactions (ADRs) were rated by two observers, using APS and ASS one year apart. The cases were ordered in a random sequence. Between-raters reliability using APS (R(est) = 0.96 and ASS (R(est) = 0.86), was very high.

4 ADR scores obtained with both methods were highly correlated ($r = 0.82$, $P < 0.001$). However, time spent using ASS was significantly longer (paired *t*-test, $t = 1.70$, $P < 0.05$).

5 These results suggest that while ASS is somewhat more complex than APS both are equally reliable and will give very similar conclusions regarding the probability of ADRs. Such algorithms must be used if the clinical assessment of ADRs is to become acceptably reliable.

Introduction

Manifestations of adverse drug reactions (ADRs) are non-specific. The suspected drug is usually administered with other drugs and frequently the adverse clinical event cannot be distinguished from the symptoms of the underlying disease. Thus, the most important problem in assessing ADRs is establishing whether there is a causal association between the suspected drug and the untoward clinical event.

Seidl *et al.* (1966) classified ADRs as having a definite, probable, possible or doubtful association to the suspected drug. However, this classification has been shown to generate large variability in between-rater assessments (Karch *et al.*, 1976; Naranjo *et al.*, 1981b). Recently, methods have been developed in an attempt to improve the reliability of the assessments of causality of ADRs (Karch & Lasagna, 1977; Kramer *et al.*, 1979; Naranjo *et al.*, 1981b). The adverse drug reactions scoring system (ASS) has been shown to generate valid and reliable assessments of ADRs (Hutchinson *et al.*, 1979). A simpler method, the adverse drug reactions probability scale (APS), developed and tested by our group yielded similar results (Naranjo *et al.*, 1981b). No comparative assessment of these procedures has been conducted as yet.

This study was undertaken to determine the correlation of the scores obtained with the ASS to those derived using the APS in rating a set of randomly selected ADRs.

Methods

Sixty-three randomly selected cases of suspected ADRs were rated independently by two raters (UB, CAN). The cases were randomized and rated in June 1978 using the APS. A year later the cases were re-ordered randomly, to minimize the influence of learning, and the raters re-analyzed the cases using the ASS. The time spent in the assessment of each ADR with both methods was recorded.

The ASS was published in 1979 and consists of a detailed algorithm that provides operational criteria for rating the probability of causation when an ADR is suspected. This algorithm provides a scoring system for six axes of decision strategy: previous general

0306–5251/82/020223–05 $01.00

experience with the drug, alternative etiologic candidates, drug levels and evidence of overdose, timing of events, dechallenge and rechallenge. The actual format for ASS is a questionnaire with 57 items. The total score range from −2 to +7 and assign the probability of the ADR (Kramer *et al.*, 1979).

The APS is a short questionnaire (10 questions) which systematically analyses the various components that must be assessed to establish a causal association between drug(s) and adverse events (i.e. pattern of response, temporal sequence, dechallenge, rechallenge, 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). Each question can be answered positive (yes), negative (no) or unknown or inapplicable (do not know) and is scored accordingly. The probability of the ADR is given by the total score and scores range from −3 to +12 (Naranjo *et al.*, 1981b).

The rating of the probability of an ADR depends on: the characteristics of the ADR; the characteristics of the rater (some raters are more reliable than others); the quality of the information (in some ADRs the information is incomplete or lacking, also it varies over time); and finally, it will also depend on the scale used to assess the ADR. Therefore, to make a proper comparison of the two scales we maintained the first three variables constant: the same raters assessed the same reactions and had identical information available. The only difference was that the reactions were assessed with the two scales (APS and ASS).

The between-raters reliability using the ASS and the APS was calculated as: (1) the product-moment correlation coefficient (r); and (2) the intra-class correlation coefficient of reliability R(est) (Spitzer *et al.*, 1978; Kramer & Feinstein, 1981), calculated as follows:

$$R(est) = \frac{S^2_s}{S^2_s + S^2_r + S^2_e}$$

Where S^2_s = variance coming from the cases; S^2_r = variance coming from the raters, and S^2_e = residual variance of error.

The R(est) combines a measure of correlation with a test in the difference of means. Therefore, this index provides a better approach to concordance because it corrects the correlation for systematic bias (Kramer & Feinstein, 1981). In addition, the scores obtained rating the 63 ADRs with the APS were correlated to those obtained when the ASS was used.

Results

Figure 1 shows that there was a high between-raters reliability when both methods were used. When the raters used the ASS the scores were highly correlated (r = 0.86, P < 0.001, **a**). Using the APS, similar results were obtained (r = 0.96, P < 0.001, **b**). This is also confirmed by the high values of the intra-class correlation coefficients of reliability (R(est) = 0.86 and 0.96, respectively).

Figure 2 shows that the scores obtained with APS were highly correlated with those obtained with ASS by both raters (RO1): r = 0.86, P < 0.001, a; and (RO2): r = 0.81, P < 0.001, b).

Figure 3 shows the correlation of the same scores when data from RO1 was combined with those of RO2. Again, the scores obtained with the APS are highly correlated to those obtained with the ASS (r = 0.82, P < 0.001).

The time (mean ± s.d.) spent using the ASS (9.52

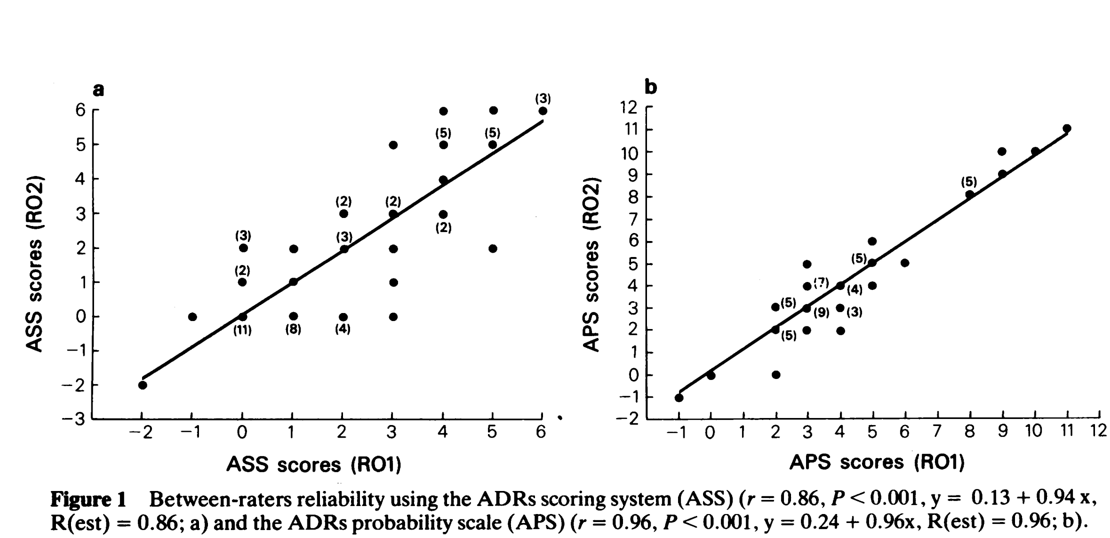

Figure 1 Between-raters reliability using the ADRs scoring system (ASS) (r = 0.86, P < 0.001, y = 0.13 + 0.94 x, R(est) = 0.86; a) and the ADRs probability scale (APS) (r = 0.96, P < 0.001, y = 0.24 + 0.96x, R(est) = 0.96; b).

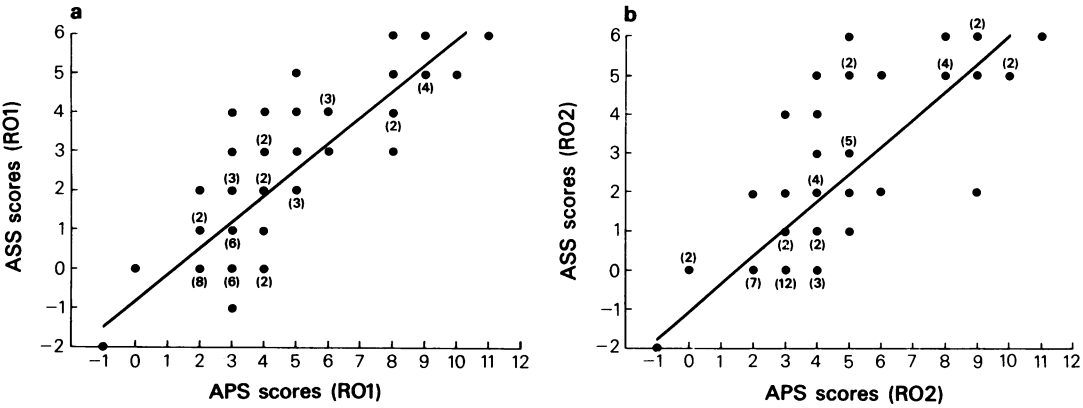

Figure 2 ADR scores using APS and ASS were highly correlated in rater RO1 (a, $r = 0.86$, $P < 0.001$, $y = -0.86 + 0.68x$) and in rater RO2 (b, $r = 0.81$, $P < 0.001$, $y = -1.02 + 0.70x$).

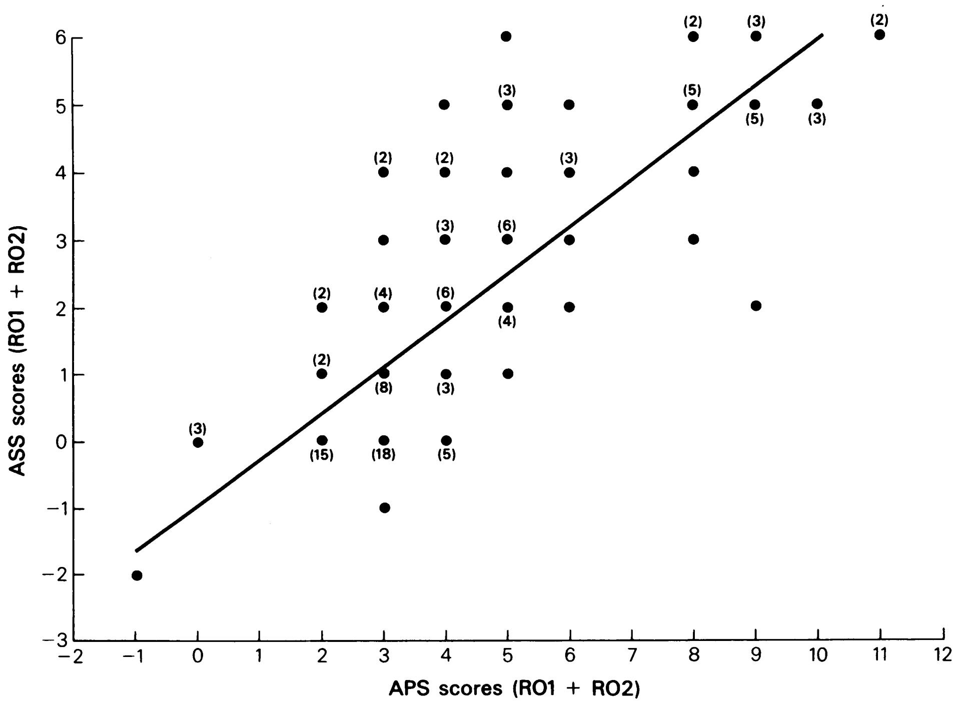

Figure 3 Combined ADR scores (RO1 + RO2) using APS and ASS were also highly correlated ($r = 0.82$, $P < 0.001$, $y = -0.94 + 0.68x$).

± 3.02 min) was slightly, but significantly longer than that using the APS (8.94 ± 3.51 min) (paired t-test = 1.70, $P < 0.05$). The recording of the time on the two occasions included the time spent reading and evaluating the case.

Discussion

This study shows that while the ADRs scoring system (ASS) is somewhat more complex than the ADRs probability scale (APS), both are equally reliable and will give similar results regarding the probability of ADRs.

Adverse drug reactions have been shown to be a major cause of morbidity (Hurwitz & Wade, 1969; Miller, 1974). Nevertheless, in all studies the diagnosis of the reactions is usually based on common sense clinical judgement. Because of the absence of clearly defined diagnostic criteria, observers frequently disagree in their assessments (Karch *et al.*, 1976; Koch-Weser *et al.*, 1977; Naranjo *et al.*, 1981b). In an attempt to develop a reproducible method for identifying ADRs, Karch & Lasagna (1977) proposed a system based on decision tables. However, no complete data about the validity and reliability of the assessments were reported (Karch & Lasagna, 1977). Since then, other methods have been developed and tested to provide a diagnostic criteria for ADRs (Kramer *et al.*, 1979; Naranjo *et al.*, 1981b). In this study we compared the assessment of ADRs obtained with two methods that have been previously shown to generate valid and reliable results (Hutchinson *et al.*, 1979; Naranjo *et al.*, 1981b).

In addition to reliability, a major requirement of a rating scale is validity. Validity means that the scores derived from the scale accurately reflect the phenomenon measured. Validity is usually assessed by comparing the results obtained with an instrument against a conventional external standard. In cases of adverse events, there is no definite standard against which to test the validity of new operational definitions of ADRs. Therefore, both methods used in this study, originally assessed validity by comparing their results against those derived from consensus of 'expert' judgements (Hutchinson *et al.*, 1979; Naranjo *et al.*, 1981b). The correlation of the scores generated by both methods is a measure of concurrent validity. Therefore, our finding that the scores were highly correlated further indicates that both methods are valid procedures that can identify which adverse events are truly ADRs.

The method proposed by Kramer *et al.* (1979) is an algorithm consisting of a maximum of 57 questions. However, despite the complex nature of the ASS, study of the instrument and experience considerably facilitates its application. Nevertheless, the time spent using ASS was slightly but significantly longer than the time spent when APS was used, even when most of the time is spent reading and evaluating the cases. Hence, the time spent answering the APS, independent of case review, is invariably much shorter than that spent on the ASS. Some could still argue that these operational definitions of ADRs are time-consuming and, therefore, they would have little application in large scale studies involving thousands of cases. In clinical practice, when the details of a case are known to the reviewer, usually less than 60 s are needed to complete the APS. In addition, the clear identification of ADRs is so important, both for the patient involved and for those who will receive the drug in the future, that some extra effort in their recognition is fully justified. The systematic application of the APS or the ASS can reduce the ambiguity that presently characterizes the assessments of ADRs.

It has been suggested that the use of algorithms for assessing ADRs may not necessarily improve reliability (Begaud *et al.*, 1980). This criticism can be interpreted as reflecting the use of an inadequate model to classify ADRs. The probability that a drug induced an ADR has been conventionally classified as definite, probable, possible or doubtful (Seidl *et al.*, 1966). However, this conventional classification assumes four discrete categories for which there is no empirical demonstration. In fact, recent studies have indicated that this categorization is an inadequate model of classification and the use of the actual ADR probability scores has been recommended (Naranjo *et al.*, 1981a). Accordingly, to prevent spurious disagreement in this study, we have only correlated the actual ADR scores obtained with both methods.

The application of operational definitions of ADRs does not replace completely clinical judgement. Knowledge of the instrument and training are essential for the successful application of the APS or ASS. Even experienced clinicians often have difficulties using operational definitions or following a structured interview (Spitzer & Endicott, 1975). There are always raters who, despite the best intentions and lengthy training, can still make unreliable assessments. Rating scales increase the precision with which certain phenomenon can be assessed and increase standardization, but cannot be expected to solve all of the problems involved on a clear definition of some clinical concepts (Spitzer & Endicott, 1975; Spitzer *et al.*, 1978). In our experience, careful study and conscientious application of either APS or ASS will yield consistently similar and reliable results by most raters. Furthermore, the results of this study indicate that when properly applied, both the APS or the ASS represent an improvement for a better assessment of ADRs.

References

BEGAUD, B., BOISSEAU, A., ALBIN, H. & DANGOUMAN, J. (1980). Comparaison de quatre methodes d'imputabilité des effets indesirables de medicaments. *Abstracts First World Conference on Clinical Pharmacology and Therapeutics*, London. Abstract No. 0771.

HUTCHINSON, T.A., LEVENTHAL, J.M., KRAMER, M.S., KARCH, F.E., LIPMAN, A.G. & FEINSTEIN, A.R. (1979). An algorithm for the operational definition of adverse drug reactions. *J. Am. med. Ass.*, **242**, 633–638.

HURWITZ, N. & WADE, O.L. (1969). Intensive hospital monitoring of adverse reactions to drugs. *Br. med. J.*, **1**, 531–536.

KARCH, F.E. & LASAGNA, L. (1975). Adverse drug reactions. A critical review. *J. Am. med. Ass.*, **234**, 1236–1241.

KARCH, F.E. & LASAGNA, L. (1977). Toward the operational definition of adverse drug reactions. *Clin. Pharmac. Ther.*, **21**, 247–254.

KARCH, F.E., SMITH, C.L., KERZNER, B., MAZULLO, J.M., WEINTRAUB, M.. & LASAGNA, L. (1976). Adverse drug reactions—a matter of opinion. *Clin. Pharmac. Ther.*, **19**, 489–492.

KRAMER, M.S. & FEINSTEIN, A.R. (1981). Clinical biostatistics LIV. The biostatistics of concordance. *Clin. Pharmac. Ther.*, **29**, 111–123.

KRAMER, M.S., LEVENTHAL, J.M., HUTCHINSON, T.A. & FEINSTEIN, A.R. (1979). An algorithm for the operational definition of adverse drug reactions. I. Background, description and instructions for use. *J. Am. med. Ass.*, **242**, 623–632.

KOCH-WESER, J., SELLERS, E.M. & ZACEST, R. (1977). The ambiguity of adverse drug reactions. *Eur. J. clin. Pharmac.*, **1**, 75–78.

MILLER, R.R. (1974). Hospital admissions due to adverse drug reactions: A report from the Boston Collaborative Drug Surveillance Program. *Arch. int. Med.*, **134**, 219–223.

NARANJO, C.A., BUSTO, U., ABEL, J.G. & SELLERS, E.M. (1981a). Empirical delineation of the probability spectrum of adverse drug reactions. *Clin. Pharmac. Ther.*, **29**, 267–268.

NARANJO, C.A., BUSTO, U., SELLERS, E.M., SANDOR, P., RUIZ, I., ROBERTS, E.A., JANECEK, E., DOMECQ, C. & GREENBLATT, D.J. (1981b). A reliable method for estimating the probability of adverse drug reactions. *Clin. Pharmac. Ther.*, **30**, 239–245.

SEIDL, L.G., THORNTON, G.F., SMITH, J.W. & CLUFF, L.E. (1966). Studies on the epidemiology of adverse drug reactions. 3. Reactions in patients on a general medical sevice. *Bull. Johns Hopkins Hosp.*, **119**, 299–315.

SPITZER, R.L. & ENDICOTT, J. (1975). Psychiatric rating scales. In *Comprehensive Textbook of Psychiatry*, Vol. 2, eds Freedman, H., Kaplan, H. & Sadock, B., pp 2015–2031. Baltimore: William & Wilkins.

SPITZER, R.L., FLEISS, J.L. & ENDICOTT, J. (1978). Problems of classification. Reliability and validity. In *Psychopharmacology: A Generation of Progress*. eds Lipton, M.A., DiMascio, A. & Killan, D.F. pp 857–869. New York: Raven Press.

(Received June 12, 1981)