A study of adverse reaction algorithms in a drug surveillance program

To improve agreement among observers, several investigators have recently proposed methods (algorithms) to standardize assessments of causality for presumed adverse drug reactions. We evaluated one such method in the context of an intensive pediatric drug surveillance program. Four observers rated 50 randomly selected case reports drawn from the program, first using only general guidelines and then, several months later, using the strict criteria of the algorithm. Agreement among observers was poor in both study phases. The presence of selected characteristics of adverse events (e.g., major severity) did not improve agreement in either phase of the study. We conclude that routine use of such algorithms in drug surveillance programs is not likely to be of benefit. (CLIN PHARMACOL THER 38:183-187, 1985.)

Carol Louik, Sc.D., Peter G. Lacouture, M.S., Allen A. Mitchell, M.D., Ralph Kauffman, M.D., Frederick H. Lovejoy, Jr., M.D., Sumner J. Yaffe, M.D., and Samuel Shapiro, M.B. *Boston, Mass., Detroit, Mich., and Bethesda, Md.*

Adverse drug reactions are a well recognized cause of morbidity, and considerable attention has been devoted to quantifying the extent of this problem. In most cases, however, the determination that a given clinical event is in fact an adverse reaction to a drug has been based on the judgment of a clinical observer. These judgments are almost certainly influenced by both the observer's experience and particular areas of concern, and it is not surprising that low rates of agreement have been documented among different investigators who evaluated selected cases of suspected adverse reactions.[5,6] In response to this problem, several researchers have proposed sets of objective criteria, or "algorithms," to be applied in a systematic manner to suspected drug reactions. These algorithms ask a series of questions, in sequence, the answers to which yield a score intended to measure the probability that the given event was in fact an adverse drug reaction.[4,7,11,12] The algorithms share certain features, and although the precise form of a particular question may differ, there are certain criteria that prove to be crucial in concluding that a clinical event is an adverse drug reaction. These include timing of the event relative to drug exposure, whether the event represents a known reaction to the drug, the possible role of the patient's condition at the time, and the effects of drug withdrawal and, where appropriate, rechallenge.

Such standardized approaches to judging adverse drug reactions are intended to produce greater consistency among evaluators, and enhanced consistency has been demonstrated when such algorithms were tested by their developers.[3,11] However, their potential utility in drug surveillance systems has not been assessed. As part of an intensive pediatric drug surveillance program, we routinely review cases of suspected adverse drug reactions and assess the likelihood that they are due to drugs. In our program, these assessments have been made informally by one observer alone. To determine whether adverse reaction algorithms enhance the consistency and usefulness of these assessments, we studied one such algorithm in the context of the Pediatric Drug Surveillance (PeDS) Program.

METHODS

The PeDS Program is an intensive drug surveillance program established at Children's Hospital in Boston to detect previously unsuspected adverse drug reactions,

From the Division of Clinical Pharmacology and Toxicology, Children's Hospital, Harvard Medical School and the Drug Epidemiology Unit, School of Public Health, Boston University School of Medicine, Boston; Children's Hospital of Michigan, Wayne State University, Detroit; and National Institute of Child Health and Human Development, Bethesda.

Supported by contracts 223-74-3196, 223-75-3036, and 1 UO1 FD 01222 from the Food and Drug Administration, Grant No. GM 24781 from the National Institute of General Medical Sciences, and Grant No. HD 17958 from the National Institute of Child Health and Human Development. The PeDS Program was initiated under a special award from the Burroughs Wellcome Fund. Dr. Louik was supported in part by National Institute of Environmental Health Sciences National Research Service Award 5T32ES07069 from the Harvard School of Public Health.

Received for publication Feb. 10, 1985; accepted April 11, 1985.

Reprint requests to: Dr. Allen A. Mitchell, Pediatric Drug Surveillance Program, Children's Hospital, 300 Longwood Ave., Boston, MA 02115.

quantify known reactions, and identify risk factors for their development. The methods have been described in detail.[9] In brief, nurse monitors are stationed on selected wards where they systematically collect data on patients admitted to these wards. Data are collected in three general areas: patient characteristics, drug exposures, and adverse events. Adverse events include any of a defined set of adverse occurrences (e.g., renal failure, convulsions, rash). Regardless of whether or not such an event is attributed to a drug by a ward observer, the nurse monitor fills out a report form that describes details of the event, date of onset, severity, treatment, duration, and outcome. In addition, reports for events believed to be drug induced include information on the identity of the suspected drug(s), dosage, and an internal assessment by the PeDS Program Director of the likelihood that the event was caused by the implicated drug. The likelihood is rated as definite, probable, possible, or doubtful.

For this investigation we selected a stratified random sample of 50 case reports from among the adverse events reported during the course of routine surveillance of patients. These included 39 "adverse reactions," stratified to include equal numbers of reactions of major, moderate, and minor severity, and 11 events not suspected to be drug induced. For phase I of the study, the 50 case reports were reviewed by three pediatric clinical pharmacologists (R. K., F. H. L., S. J. Y.) who were not directly involved with the PeDS Program and by the PeDS Program Director (A. A. M.), who routinely assesses adverse reaction reports in the Program. To mimic the way assessments are made within the PeDS Program, all reviewers were asked to make an informal judgment as to the probability that each case was an adverse reaction to a suspected drug and to rate each case as "definite," "probable," "possible," "doubtful," or "insufficient information." These terms were deliberately left undefined; we relied on the intuitive understanding of each reviewer to replicate the manner in which these assessments have traditionally been made. The consultant reviewers were told only that we were evaluating our assessments of adverse reactions; at this time no mention was made of algorithms.

Phase II of the study began 3 months after the completion of the first set of reviews. Our consultants were again sent the same 50 case reports, but this time they were asked to assess them using the algorithm proposed by Naranjo et al.[11] This algorithm was chosen because it incorporates the principal features of most others and was judged in our pilot study to be easier to learn and faster to use. This algorithm consists of a set of 10 questions, each of which can be answered "yes," "no," or "do not know," and each of the 10 responses is assigned a score. Our reviewers were asked to answer each question; the scores were tallied and each case report was then categorized according to the established criteria as a "definite," "probable," "possible," or "doubtful" drug-induced event.[11]

In both phases I and II, differences between raters were assessed with the test for marginal homogeneity proposed by Mantel and Byar.[8] This is a generalization of the Mantel-Haenszel test for matched samples. We used Cohen's[1] κ to measure pairwise agreement between all pairs of raters, and then used an approach suggested by Fleiss[2] to measure agreement among more than two raters by taking the average of all pairwise κ values. The statistic κ is a measure of the extent of agreement beyond that expected by chance; it ranges in value from -1 to $+1$. Values >0.75 are generally considered to represent "good" agreement, values between 0.4 and 0.75 to represent "fair" agreement, and values <0.4 to represent "poor" agreement. We also measured agreement with the intraclass correlation coefficient (ρ), a statistic that is essentially equivalent to a weighted κ provided the weights have a particular form.[2] Unlike the unweighted κ, the intraclass correlation coefficient gives partial weighting to partial agreement (e.g., ratings from two reviewers of "definite" and "probable" receive more credit than ratings of "definite" and "doubtful") and would therefore be expected to give slightly higher values than unweighted κ when used to measure agreement in ordered categories.

RESULTS

Table I shows the distribution of ratings for each rater. Because "insufficient information" was used only five times, these ratings were reclassified as "possible" for all analyses. The Mantel-Byar tests for marginal homogeneity indicate that in both phases of the study there were substantial differences in the frequency distributions of ratings for each rater. Thus in the aggregate, some raters were considerably more likely to choose a "definite" or "doubtful" rating. For example, in phase I, rater 2 used "definite" only four times (2%), while rater 1 chose it 15 times (30%). This difference was somewhat less marked in phase II, but still statistically significant.

In Table II, agreement in assignment of ratings to individual cases as measured by both the K statistic and the intraclass correlation coefficient is shown for all pairwise combinations of raters as well as for all four raters considered together. As expected, in all cases the

Table I. Distribution of assessments of 50 adverse events according to rater

	Phase I				*Phase II*			
Rater	*Definite*	*Probable*	*Possible*	*Doubtful*	*Definite*	*Probable*	*Possible*	*Doubtful*
1	15	10	16	9	0	24	25	1
2	4	20	19	7	5	10	32	3
3	13	12	11	14	1	17	31	1
4	12	13	6	19	7	14	27	2
	X^2 (df = 9; marginal homogeneity) = 35.4 (P < 0.005).				X^2 (df = 9; marginal homogeneity) = 21.7 (P < 0.01).			

Table II. Agreement among raters of 50 adverse events

	Phase I		*Phase II*	
Sets of raters	K	*r*	K	*r*
1-2	0.12	0.40	0.14	0.30
1-3	0.44	0.74	0.42	0.51
1-4	0.22	0.38	0.31	0.50
2-3	0.21	0.40	0.20	0.38
2-4	0.10	0.28	0.24	0.58
3-4	0.27	0.56	0.36	0.62
All raters	0.24	0.51	0.30	0.48

intraclass correlation coefficient indicates a greater degree of agreement than K. In phase I, when assessments were made without specified guidelines, agreement was poor as measured by either statistic, except for one pair of raters for whom r indicates fairly good agreement (r = 0.74). Application of the algorithm to assess causality in phase II of the study did not improve agreement. Although the overall K indicates slightly better agreement, r is quite similar for the phase I and phase II assessments, and pairwise comparisons show no trend in either direction.

We stratified the events according to characteristics that might affect the level of agreement. Factors selected were severity of the event, whether the event was originally reported as drug attributed, whether treatment was required, and whether blood levels of the suspected drug had been measured (Table III). In general, for a given event characteristic, measures of agreement did not vary substantially in either phase. For example, in phase I, events that were of major severity were no more likely to yield agreement among raters than were events of minor severity; this finding held for phase II as well. However, agreement was better in both phases among the 39 events originally attributed to drugs than among the 11 events that were not.

When results from phase I were compared with those from phase II, there was little evidence that use of the algorithm improved agreement for most of the event characteristics studied, although phase II agreement measures were better for those case reports that included a blood level value of the suspected drug. However, in no category did use of the algorithm raise agreement to "good" levels.

DISCUSSION

For any individual adverse event it is rarely possible to know with absolute certainty whether the event is drug induced. As a first step in approaching truth, one would hope to achieve a consensus among experts. Such consensus is difficult to obtain when judgments are made intuitively and when unspecified criteria are used to make the determination of drug involvement.[5,6] This problem led several investigators to develop methods of standardizing assessments, and reports have suggested that agreement is indeed improved when these methods are used. In particular, Naranjo et al.[11] used a study design similar to ours and demonstrated a change in the intraclass correlation coefficient from 0.49 ("fair") to 0.92 ("good").

Our study is the first to consider the utility of these algorithms in the assessment of adverse events routinely identified as part of an intensive drug surveillance pro-

Table III. Agreement among raters according to selected characteristics of events

	No. of cases	*Phase I*		*Phase II*	
		K	*r*	K	*r*
Severity					
Major	14	0.16	0.50	0.26	0.50
Moderate	16	0.21	0.43	0.30	0.53
Minor	20	0.27	0.53	0.31	0.42
Drug attribution					
Yes	39	0.18	0.36	0.24	0.43
No	11	0.05	0.16	−0.20	−0.17
Treatment required					
Yes	24	0.22	0.58	0.34	0.51
No	26	0.24	0.42	0.26	0.48
Blood levels measured					
Yes	12	0.15	0.25	0.28	0.56
No	38	0.24	0.52	0.24	0.29

gram. While levels of agreement before use of the algorithm were similar in both the study of Naranjo et al.[11] ($r = 0.49$) and our own ($r = 0.51$), we were unable to demonstrate that application of the algorithm led to improved agreement among raters.

We believe there are two factors that account for our discordant findings. First, our raters were pediatric clinical pharmacologists who were not experienced in the use of any of the proposed algorithms. This criterion for selection of the raters was intended to assure that their first "informal" assessments would not be influenced by their unconscious application of any standardized set of criteria. If this explanation is correct, it suggests that simple application of these algorithms as they are published will not automatically lead to greater consensus.

Second, our case reports differ from those used in other studies to test algorithms. While we used reports derived from our routine surveillance, Naranjo et al. used published case reports. Unlike reports from intensive surveillance programs, which are necessarily brief and often based on incomplete information, published reports are more complete—in fact, they must contain sufficient evidence of a cause and effect relationship to justify publication. Thus such reports are more likely to include information on many of the important components required by the Naranjo et al. (and other) algorithms. For example, blood levels of the suspected drug would almost certainly be included in published reports, but may only infrequently be available for suspected reactions observed by a surveillance program. It is interesting that in our study, the most marked improvement in agreement with the use of an algorithm occurred for those events for which blood levels were available.

The role of adverse drug reaction algorithms in a drug surveillance program must be kept in perspective. These algorithms are not designed to identify previously unreported adverse drug reactions, and their inclusion in a drug surveillance program would not enhance the capacity to detect such previously unrecognized reactions. That objective is better met by the use of "event surveillance," in which attributions of causality are not relevant.[10] Because event surveillance applies to populations of subjects, it cannot provide useful assessments of causality for individual cases. Algorithms would seem to be an alternative approach to this problem. However, our findings indicate that the use of an algorithm in an intensive drug surveillance program does not improve the uniformity (or validity) of adverse drug reaction assessments.

We thank Jane Sheehan, R.N., Project Coordinator and the nurse monitors of the PeDS Program who collected patient data, and John Farrell and Leonard Gaetano for computer programming assistance.

References

1. Cohen J: A coefficient of agreement for nominal scales. Educ Psychol Meas **20:**37-46, 1960.
2. Fleiss JL: Measuring nominal scale agreement among many raters. Psych Bull **76:**378-382, 1971.
3. 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.
4. Karch FE, Lasagna L: Toward the operational identifi-

cation of adverse drug reactions. CLIN PHARMACOL THER **21**:247-254, 1977.

5. Karch FE, Smith CL, Kerzner B, et al: Adverse drug reactions: A matter of opinion. CLIN PHARMACOL THER **19**:489-492, 1976.
6. Koch-Weser J, Sellers EM, Zacest R: The ambiguity of adverse drug reactions. Eur J Clin Pharmacol **1**:75-78, 1977.
7. 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.
8. Mantel N, Byar DP: Marginal homogeneity, symmetry, and independence. Commun Statist Theor Math A7:953-976, 1978.
9. Mitchell AA, Goldman P, Shapiro S, Slone D: Drug utilization and reported adverse reactions in hospitalized children. Am J Epidemiol **110**:196-204, 1979.
10. Mitchell AA, Slone D, Shapiro S, Goldman P: Adverse drug effects and drug surveillance. *In* Yaffe SJ, editor: Pediatric pharmacology: Therapeutic principles in practice. New York, 1980, Grune & Stratton Inc., pp 65-78.
11. Naranjo CA, Busto U, Sellers EM, et al: A method for estimating the probability of adverse drug reactions. CLIN PHARMACOL THER **30**:239-245, 1981.
12. Venulet J, Ciucii A, Berneker GC: Standardized assessment of drug-adverse reaction associations—Rationale and experience. Int J Clin Pharmacol **18**:381-388, 1980.