A Critical Appraisal of Toxicity Indexes in Rheumatology

PAUL M. PELOSO, JAMES G. WRIGHT, and CLAIRE BOMBARDIER

ABSTRACT. Our objective was to design a grading system to judge the methodological quality of toxicity symptoms assessment in rheumatologic randomized controlled trials, and to use this grading system to evaluate rheumatology case report forms. For comparison, we also evaluated instruments derived in other clinical specialties. To further determine whether variability seen on the rheumatology case report forms would lead to variability of side effect measurement in a clinical trial setting, we conducted a survey of investigators and their coordinators in a multicenter randomized trial. We used Feinstein's definition of sensibility and a conceptual model of the process of toxicity measurement in the clinical trial setting to develop a grading system. Thirty case report forms were obtained from a convenience sample, and 7 other instruments from other disciplines were obtained through literature review. These forms were evaluated in a blinded manner by 3 reviewers. A survey of 30 trial staff from 15 centers involved in a multicenter, randomized trial of 2 active nonsteroidal antiinflammatory drugs (NSAID) in rheumatoid arthritis (RA) was undertaken to evaluate whether the variability identified in the review of the case report forms translated into variability in the measurement of side effects in clinical trial setting. The rheumatology case report forms had a mean score of 9.1 (SD = 2.6) out of 20, where 0 is the worst score and 20 is a perfect score. The structured instruments from other disciplines had a mean score of 14.9 (SD = 2.7). These means were significantly different, with $p = 0.001$. The reviewers showed good levels of agreement, with kappa values of 0.93, 0.89, and 0.80 for intrarater agreement, and 0.70 (95% CI 0.58 to 0.82) for interrater agreement. The survey demonstrated variability in the manner of enquiry and recording of side effects in the case report form. We conclude the current case report forms in rheumatology failed to meet half the defined criteria expected of a valid instrument, while structured forms from other disciplines performed much better. These identified weaknesses lead to variability in side effect measurement in the usual clinic setting, as demonstrated by a survey in a multicenter randomized trial. (*J Rheumatol 1995;22:989–94*)

Key Indexing Terms:
TOXICITY CASE REPORT FORMS QUALITY

The decision to prescribe a drug, device, or program to a patient depends on knowledge of both the expected benefit and the expected side effects. Faced with 2 therapies of similar efficacy, most clinicians would choose the agent that is least toxic, provided cost and convenience were also similar. Until recently, there has been little consensus in rheumatology on measuring either efficacy or toxicity. The clinician might ask, "What represents a clinically important change from the baseline state, and what represents a clinically important toxicity, and how should these 2 be measured?"

From the Clinical Epidemiology Division, Wellesley Research Institute, and the Rheumatic Diseases Unit, Wellesley Hospital, the Department of Surgery, Hospital for Sick Children, and the Department of Medicine, University of Toronto, Toronto, ON, Canada.

Dr. Peloso is supported by a Fellowship from The Arthritis Society; Dr. Wright is supported by a grant from the Ontario Ministry of Health.

P.M. Peloso, MD, FRCPC, MSc, Clinical Research Fellow, Wellesley Hospital and the University of Toronto; J.G. Wright, MD, FRCPS, MPH, Assistant Professor, Department of Pediatrics and Orthopedics, Hospital for Sick Children; C. Bombardier, Professor of Medicine, Rheumatic Diseases Unit, Wellesley Hospital and the Department of Medicine, University of Toronto.

Address reprint requests to Dr. P.M. Peloso, RDU, 3rd Floor Ellis Hall, Royal University Hospital, Saskatoon, SK, Canada S7N 0W8.

Recent work by the American College of Rheumatology has focussed on the standardization of outcome measures in rheumatology. Most of this work has centered on efficacy measures[1,2]. Much less attention has been devoted to the other side of the equation — measuring side effects. Both are important to the clinical decision making process, and both need to be uniformly, precisely and reliably measured. This is particularly important in the rheumatic diseases, where physicians treat chronic disorders over decades, using a variety of drugs known to have narrow efficacy – toxicity windows.

There are several distinctions to measuring drug toxicity that must be appreciated — both immediate and delayed effects can be studied, as can patient symptoms, physical signs, and laboratory abnormalities. These distinctions are represented in Figure 1. Figure 1 also shows the different stages of drug development. Each of these stages and types of toxicities may require different instruments, personnel, or study designs to be properly measured[3-5].

Our purpose was to evaluate the process of measuring patient symptoms in the rheumatology clinical trial specifically. It has been shown in other disciplines that the measurement of patient symptoms is an important predictor of both patient clinical outcomes[6], and health related quality of

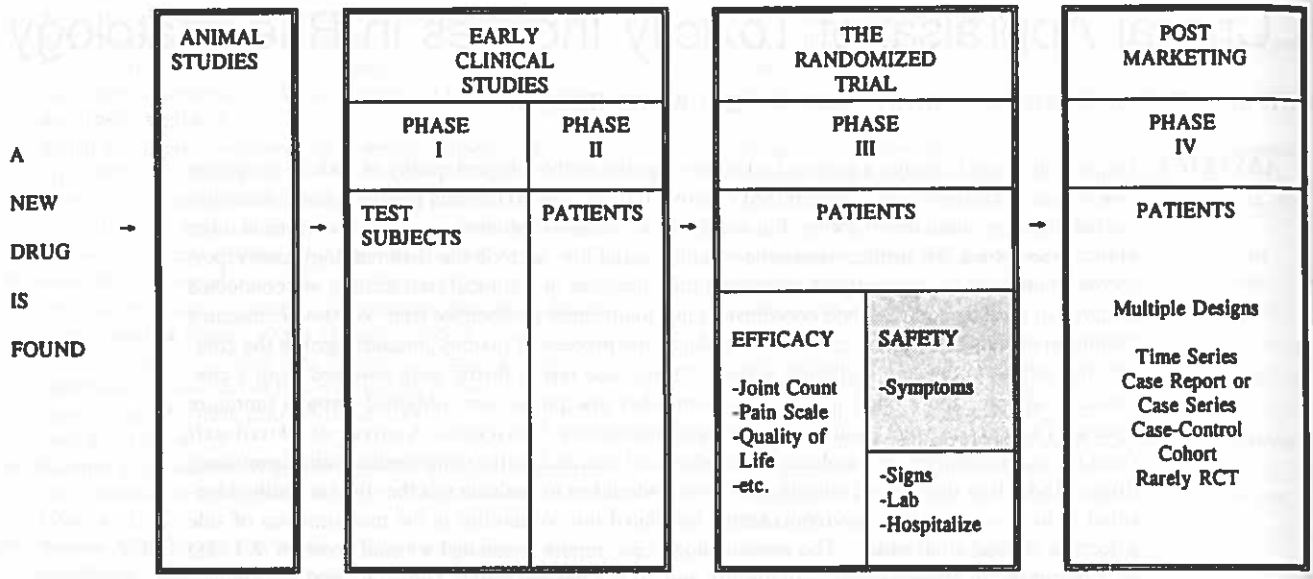

Fig. 1. The process of drug evaluation. There are many phases in the evaluation of toxicity. The shaded area indicates the focus of this paper — symptom measurement in the randomized controlled trial.

life[7]. It is likely that such a relationship holds for rheumatic patients as well. The case report forms used in the rheumatologic setting can be contrasted to the forms used in other disciplines, where greater efforts have been made to standardize symptom collection.

We developed a grading form based on Feinstein's definitions of sensibility[8] (grading form shown as Appendix), and graded a convenience sample of 30 rheumatology case report forms, as well as 7 forms derived from other disciplines[7-15]. We then examined for the implications of any inconsistencies and omissions on these forms by surveying physicians and their coordinators about the procedures used to measure patient symptoms in patients with RA who had participated in a multicenter randomized controlled trial.

MATERIALS AND METHODS

Using Feinstein's definition of sensibility[8] and a conceptual model of potential steps and interactions in the rheumatology randomized controlled trial, we developed a scale of items that measured the methodologic quality of the case report forms and other indexes from a rheumatologic point of view. The schematic model is shown in Figure 2. The rating scale was developed in a blinded fashion, without indepth knowledge of any one instrument, and was pretested by all 3 authors in 3 iterative steps designed to eliminate inconsistent, repetitive, or inapplicable items. A glossary of terms and definitions for each item was developed and agreed upon by the 3 evaluators before the final evaluation of the toxicity forms. The final version of the evaluation criteria was derived from the authors' opinions about the most relevant 20 items. One item was awarded as a bonus point, and was not part of the 20 core scoring items. All evaluations of the case report forms were blinded to the drugs under study, sponsoring pharmaceutical company, investigator, or year of use, in both the training sessions and the final scoring session. All disagreements were resolved by a consensus approach, once it was determined that the disagreements were not related to simple oversights.

The 30 case report forms obtained were a convenience sample, derived from randomized trials in RA conducted at Wellesley Hospital during the

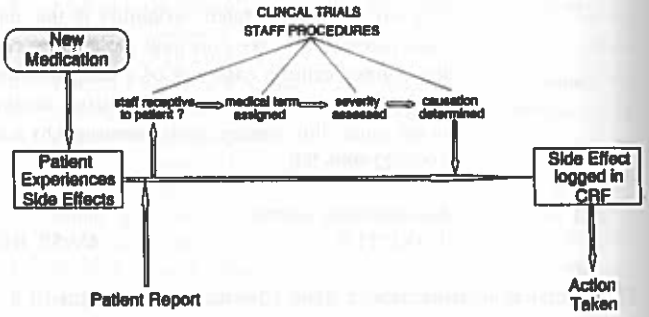

Fig. 2. Steps for reporting toxicity in the randomized controlled trial (CRF = case report form).

years 1984–92 (n = 16), and from other North American RA investigators (n = 14). The case report forms evaluated are available from the authors on request. All case report forms came from published trials. The review and grading of the protocol was not part of the standardized sensibility review process, as obtaining a complete sample of protocols was not possible. The nonrheumatology toxicity indexes were identified through a literature review using *Medline* and contact with investigators in other disciplines. All nonrheumatologic instruments had been used in drug trials in the particular disciplines.

The survey was designed to test the clinical relevance of the deficiencies in toxicity case report forms. It was conducted at the final investigators' meeting of a Canadian multicenter trial comparing 2 active NSAID in the treatment of RA. The physicians were surveyed independently of the coordinators, and each group had no knowledge of the others' answers. The investigators had met on 2 occasions before trial commencement to discuss the choice and the measurement of outcome variables. Each center was also trained in the performance of a standardized joint count and the trial quality of life measure, the Patient Elicitation Technique (PET)[16,17]. However, apart from instructions in the protocol, there were no instructions on how to standardize symptom measurement for toxicity assessments.

Statistical analysis. Data from the quality analysis and the survey were entered into Epi-Info[18] and analyzed using PC-Agree[19] and SAS version 6.03[20]. The agreement among and within assessors was evaluated using the

kappa statistic[21,22]. This statistic evaluates agreement beyond that which would occur through chance alone. The scores between the rheumatology forms and the other disciplines were contrasted using the modified T test for unpaired data and the ranked sum test, with ≤ 0.05, 2 tailed, considered statistically significant.

RESULTS

The 30 case report forms used to perform the quality review consisted of 3 auranofin trials, 3 azathioprine trials, 3 trials of biologics, one each from cyclosporine and deflazacort, 2 from hydroxychloroquine, 10 of methotrexate, 5 of myochrysine, and 2 of sulfasalazine. The year of study completion ranged from 1984 to 1992. For 18 of 30 case report forms studied, the pharmaceutical industry was the primary motivation for the study, determined by a statement to that effect in the protocol, a company logo on the case report form, or by communication with the individual having supplied the case report form. The 7 nonrheumatology based forms came from oncology (3 forms), psychiatry (2 forms), and general medicine (2 forms).

The final grades for the toxicity instruments are shown in Table 1. The nonrheumatology forms had higher scores, with a mean score of 14.9/20, a range of 11–19, and a median of 14, than the rheumatology based case report forms, where the mean score was 9.1/20, the range 6–15, and the median 8 ($p = 0.001$). The year of trial completion for the rheumatology case report form was a predictor of the quality score on linear regression, with $R^2 = 0.13$ and $p = 0.051$, with later years having higher scores. Pharmaceutical sponsorship was not a significant predictor of the final grade in this sample.

Agreement on the rating was very good within individual raters, with kappa values of 0.93, 0.89, and 0.80. Kappa values range from −1.0 to 1.0; values above 0.8 are considered excellent, and values below 0.2 poor. The kappa value for the overall agreement among reviewers was 0.70, with 95% confidence limits of 0.58–0.92, indicating very good agreement among reviewers[21].

Clinical significance. We wondered whether the variability found in the case report forms would translate into variability in the clinical setting. To investigate this, we surveyed the physicians and their study coordinators involved in a Canadian multicenter, randomized, double blinded trial comparing 2 active NSAID in the treatment of RA. Fifteen centers were surveyed, with 28 of the 30 study personnel providing usable responses.

In response to the question "What Question was used to Enquire about Side Effects in the Trial," 3 categories of responses were obtained: those directed toward side effects, such as "Any problems with the medication?" or "Any side effects since last visit?", which require some judgments about attribution from the patients (21/28); those that directed the questions toward symptoms, such as "How have you been feeling in the last 2 weeks?" and "Have you had any different symptoms since last visit)", which do not require the patient to make any inferences about causation (5/28); and those questions that enquired about the general health of the patient, such as "Any changes, good or bad, in your condition?" (2/28).

In response to the question "When were side effects discussed?", 50% stated it was at the beginning of the visit, 33% stated any time, and 17% stated at the end of the visit. The issues related to side effects were discussed before the joint count 75% of the time, after the joint count 21% of the time, and at no set point 4% of the time. The response to the question"What was recorded in the case report form?" disclosed that 75% of the time, anything reported by the patient as an unwanted experience was recorded, 21% of the time, the physician had to think that the drug might at least have caused the symptom, and 4% of the time, the patient had to feel there was an association between the drug and

Table 1. *A comparison of quality scores for rheumatology case report forms versus nonrheumatology based indexes*

Form	Administered By	Discipline	Quality Score (maximum = 20)	Reference
30 Case report forms	Study staff	Rheumatology	Mean 9.1 SD = 2.7 Median = 8 Min-Max = 6–15	
UKU, EPS	Patient or staff, staff	Psychiatry	13, 11	11, 12
WHO, MDA, ECOG	Staff	Oncology	15, 14, 14	9, 10, 14
Diabetes	Patient	General medicine	18	8
POSI	Patient	General medicine	19	7
All nonrheumatology instruments in total				Mean = 14.9 SD = 2.7 Median = 14

The mean values for rheumatology versus nonrheumatology forms are statistically significantly different, using a modified 2 tailed T test, at $p = 0.001$. UKU = European Psychiatry Side Effect Questionnaire; EPS = Extra Pyramidal Side Effect Questionnaire; MDA = M.D. Anderson Oncology Questionnaire; ECOG = Eastern Cooperative Oncology Group Questionnaire; POSI = Patient Oriented Symptoms Index.

the symptom, in order for the symptom to appear in the case report form. These results suggested that the lack of clear instruction on the case report form did lead to differences in how symptoms were elicited, when they were elicited, and what was recorded in the case report form.

DISCUSSION

We have shown that standardized criteria for the judgment of case report forms as a toxicity measurement tool can be developed from a knowledge of Feinstein's sensibility criteria[8] and from a conceptual model of side effect measurement in the rheumatology clinic. These criteria can then be applied in a reliable way to case report forms and other toxicity indexes. Such an analysis revealed that the unstructured case report form does not perform as well as more structured forms developed by other disciplines to measure toxicity, even when our criteria are biased toward the rheumatologic setting.

Since we examined the actual case report forms primarily, it might be suggested that the categories that scored poorly on our case report form review were present in the protocol, and that we have judged the rheumatology case report forms harshly. We addressed this concern by reviewing the protocols of 15 of the 30 case report forms, looking for information missing on the case report form but present in the protocol. In only 2 instances did the protocol provide any additional information. In both those instances, the additional information was a defined question on how to probe for side effects. When the analysis was redone accounting for the possibility that some case report forms were scored lower by 1 or 2 points on the initial evaluation, the overall conclusions were unchanged. That is, the structured forms from other disciplines still outperformed the rheumatology case report forms. It seems logical that if instructions are not present on the case report form, or if the case report form is not clear on where to find such instructions in the protocol, a consistently applied approach to measuring toxicity is unlikely, even if such instructions do exist. In centers where several protocols are concurrently under study, the risk of deviating from any one protocol would seem to be higher.

The case report forms failed to meet over half the criteria one might expect from a methodologically valid instrument in this blinded assessment. The major deficiencies identified in the review related to a lack of clear instructions on who should elicit toxicity, how it should be elicited, what time frame should be used to elicit toxicity, and how judgments should be made about both the severity of the side effect and the likelihood of the drug having caused the side effect. The main advantages of the usual rheumatology case report form are that it is familiar, easy to use, and acceptable to the regulatory bodies approving new drug use.

Our rating criteria were derived from a knowledge of the process of side effect collection in the rheumatology trial setting, and where inconsistencies might occur, and from Feinstein's sensibility criteria for the evaluation of indexes. Earlier versions included more items, but we felt that this list of 20 represented the most important items. Further work will be needed to determine whether important items have been omitted, and whether some items should preferentially be given greater weight. Our scoring form allows a starting point for discussion, however.

Since the use of structured instruments in other disciplines is not universally practised, it might be suggested that the comparison of the case report forms from rheumatology to the structured instruments is unfair. Perhaps the rheumatological case report forms would not have looked as bad in comparison to case report forms in oncology and psychiatry. Work in oncology suggests that only 40% of oncology prospective trials after the year 1990 used standardized instruments to measure toxicity, although appeals for greater use of such standardized instruments are being made[23]. The lack of use of more methodologically sound instruments should lead logically to a discussion of how to encourage their use. Until recently, the use of quality of life instruments was not common in rheumatology trials, but quality of life measurements have been shown to be valuable tools in comparisons of different drugs in one disease, or comparisons across diseases[16].

One potential barrier to the use of a standard instrument for toxicity might be the argument that the types of toxicities vary too widely in rheumatology to make a single instrument practical for all the drugs used to treat RA. There are, however, a limited number of ways that the questions about symptoms can be asked, and the majority of the structured instruments are remarkable for their similarities. These checklists commonly include 35–50 questions completed by the patient, and require 5 to 10 min to complete[11,12,24]. The numerical values derived from these checklists can be used to evaluate the effect of drugs on quality of life[7], or provide a summary score of toxicities experienced in a clinical trial.

The Fries Toxicity Index[24] is a scoring index that allows comparisons across different drugs. The index does not include a standard data collection instrument to elicit symptoms of toxicity in the randomized controlled trial, but it does provide a weighting scheme for symptom importance, and a summary score. The scoring scheme has been derived from the long version of the Health Assessment Questionnaire administered to subjects every 6 months as part of the Arthritis, Rheumatism and Aging Medical Information System (ARAMIS) database. Any standardized symptom index used in rheumatology trials could take advantage of this index to provide a summary of toxicity for the trial. The summaries derived from the Fries Toxicity Index would clearly facilitate comparisons of drug toxicities in clinical trials, and it has already proved clinically useful in summarizing toxicity in longterm cohorts of patients with RA[25] and in metaanalytic summaries of drug toxicities in RA[26].

It would seem that deficiencies identified in the case report forms do translate into meaningful clinical differences among users. The results of the survey of a group of Canadian rheumatologists suggest that there are indeed differences in how various aspects of toxicity measurement are carried out. Whether this survey is generalizable cannot be stated with certainty, but this was an experienced group of trialists. Whether this survey captured accurately what occurred in the clinic could only be determined with certainty by some blinded methods. Such methods might include video or audio taping the encounters with patients, or having neutral observers rate the encounters related to toxicity, with the investigators blinded to the purpose of the observer.

CONCLUSION

It would seem that current rheumatology case report forms do not meet the majority of the criteria one might expect of a methodologically sound instrument. This leads to variability in the clinical setting relating to toxicity measurement. Other disciplines have designed instruments to circumvent some of the problems noted in rheumatology case report forms. Toxicity measurement in rheumatology would be greatly enhanced by a standardized approach to eliciting symptoms, an agreed upon core set of toxicities to be measured across all trials, and a uniform structure for the rheumatology case report form, with clear definitions for any judgments required of the clinical investigator in rheumatology.

APPENDIX: Scoring for Sensibility Rating. Final score is 20 points (one bonus point possible)

	Score
Purpose	
1. Is the intended population specified?	Yes = 1
Replicability	
2a. Are all the elements of the index provided in the notes/report?	Yes = 1
2b. (i) Are there instructions for all items?	Yes = 1
2b. (ii) Are these instructions clear?	Yes = 1
2c. Does the index state who should elicit toxicity?	Yes = 1
2d. Are written instructions for side effect elicitation given?	Yes = 1
2e. Are regular elicitation intervals defined on the form?	Yes = 1
Comprehensiveness and discrimination	
3a. (i) Does the index allow for extra side effects to be reported?	Yes = 1
3a. (ii) Is the index too narrowly focussed?	No = 1
3b. Does the index provide gradations for severity?	Yes = 1
3b. (i) Are the grades clearly defined?	Yes = 1
3b. (ii) Do these definitions make sense?	Yes = 1
3c. (i) Are the categories or types mutually exclusive?	Yes = 1
3c. (ii) Are the categories all inclusive?	Yes = 1
3d. (ii) Are instructions for assessing causation given?	Yes = 1
3e. Are paraclinical data defined appropriately?	Yes = 1 (bonus)
Face validity	
4a. Are the questions posed in a neutral manner?	Yes = 1
4b. Have important items been omitted?	No = 1
4c. Is this index appropriate for rheumatology?	Yes = 1
4d. Does the index have a summary score?	Yes = 1
Ease of use	
5. Would the target audience find the scale easy to use?	Yes = 1

ACKNOWLEDGMENT

We thank Dr. Charles Goldsmith for review of this manuscript and statistical advice, and Dr. D. Felson, Dr. H.J. Williams, Dr. M.D. Brundage, and Dr. M. Keating, who supplied case report forms and information about other indexes.

REFERENCES

1. Goldsmith CH, Boers M, Bombardier C, Tugwell P, for the OMERACT Committee: Criteria for clinically important change in outcomes:. Development, scoring and evaluation in rheumatoid arthritis patient and trial profiles. *J Rheumatol 1993;*20:561–5.
2. Felson D, Anderson JJ, Boers M, *et al:* The ACR preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis Rheum 1993;*36:729–39.
3. Grahame-Smith DG, Aronson JK: Adverse reactions to drugs. In: *The Oxford Textbook of Clinical Pharmacology.* Oxford: Oxford University Press, 1984:132–57.
4. Sackett DL, Haynes RB, Tugwell P, Guyatt GH: *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Toronto: Little, Brown and Company, 1991.
5. Stolley PD, Strom BL: Evaluating and monitoring the safety and efficacy of drug therapy and surgery. *J Chron Dis 1986;*39:1145–55.
6. Awad AG, Hogan TP: Subjective response to neuroleptic and the quality of life: Implications for treatment outcome. *Acta Psychiatr Scand 1994;*in press.
7. Testa MA, Anderson RB, Nackley JF, Hollenberg NK: Quality of life and antihypertensive therapy in men. *New Engl J Med 1993;*328:907–13.
8. Feinstein AR: *Clinimetrics.* New Haven: Yale University Press, 1987.
9. Levine M: The Patient Oriented Symptoms Index — the POSI. Unpublished manuscript. Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada.
10. Testa MA, Simonson DC: Measuring quality of life in hypertensive patients with diabetes. *Postgrad Med J 1988;*64(suppl 3):50–8.
11. Oken MM, Creech RH, Tormey DC, *et al:* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol 1982;*5:649–55.
12. WHO Handbook for Reporting Results of Cancer Treatment. Geneva: World Health Organization, 1979.
13. Ajani JR, Welch SR, Raber NR, Fields WS, Krakoff IH: Comprehensive criteria for assessing therapy-induced toxicity. *Cancer Invest 1990;*8:147–59.
14. Lingjaerde O, Ahlfors UG, Bech P, Dencker SJ, Elgen K: The UKU Side Effects Rating Scale: A scale for the registration of unwanted effects of psychotropics. *Acta Psychiatr Scand 1986;*334:81–100.
15. Lingjaerde O, Ahlfors UG, Bech P, Dencker SJ, Elgen K: The UKU Side Effect Study: Aims and methods. *Acta Psychiatr Scand 1987;*334:29–79.
16. Deyo RA, Patrick DL: Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Med Care 1989;*27(suppl 3):S254–68.
17. Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life, and utility in rheumatoid arthritis. *Arthritis Rheum 1990;*33:591–601.
18. Dean AG, Dean JA, Dicker RC: *Epi-Info, Version 5: A Word Processing, Database and Statistics Program for Epidemiology on Micro-computers.* Stone Mountain, GA: USD Inc., 1990.
19. Walter SD, Cook RJ: *PC-Agree Version 3.0: A PC Program*

for the Analysis of Interobserver Variation. Hamilton, ON: Department of Clinical Epidemiology, McMaster University, 1991.
20. Freeman GC, Godfrey KW: *SAS: Release 6.03 Edition.* Cary, NC: SAS Institute Inc., 1988.
21. Landis JR, Koch GG: The measurement of interobserver agreement for categorical data. *Biometrics 1977;*33A:159–74.
22. Kramer MS, Feinstein AR: Clinical Biostatistics LIV. The Biostatistics of Concordance. *Clin Pharmacol Ther 1981;*29:111–23.
23. Pater J, Brundage M, *et al:* Outcome Measurement in Oncology. *Ann Royal Coll Phys Surg Canada 1994;*27:174–8.
24. Fries JF, Spitz PW, Williams CA, Bloch DA, Singh G, Hubert HB: A toxicity index for comparison of side effects among different drugs. *Arthritis Rheum 1990;*33:121–30.
25. Fries JF, Williams CA, Raney D, Bloch DA: The relative toxicity of disease modifying anti-rheumatic drugs. *Arthritis Rheum 1990;*36:297–306.
26. Felson DT, Anderson JJ, Meenan RF: Use of short-term efficacy/toxicity tradeoffs to select second-line drugs in rheumatoid arthritis. *Arthritis Rheum 1992;*35:1117–25.

ARAMIS and Toxicity Measurement

JAMES F. FRIES

ABSTRACT. Side effects of medications make up an important part of adverse outcomes experienced by patients with rheumatic diseases. Quantitative measures to assess toxicity, however, have not been available, and this lack has limited estimates of the magnitude of effects and of differences in side effects among different drugs. This paper describes the development of the Arthritis, Rheumatism and Aging Medical Information System (ARAMIS) Toxicity Index and the issues arising in construction of such an index, and reviews early results in comparing toxicities of antirheumatic drugs. Findings have had major value in revising therapeutic strategies for rheumatic diseases, particularly rheumatoid arthritis, and have set the stage for development of toxicity–therapeutic ratios for different drugs. (*J Rheumatol 1995;22:995–7*)

Key Indexing Terms:
DRUG TOXICITY TOXICITY INDEX ARAMIS POSTMARKETING SURVEILLANCE

Rational clinical decision making, perhaps almost an oxymoron, requires quantification of the unquantifiable. If the typical decision requires selection of drug A or drug B, then the decision maker requires, at a minimum, an accurate knowledge of all the good things that can be expected to happen with each drug, and similar knowledge of all of the bad things that can be expected to follow each treatment. In the most profound sense this knowledge is required not just for the average of a group of individuals but for the specific patient for whom the decision looms.

To approach the issue of comparative effectiveness or of comparative toxicity of 2 agents, indexes must be constructed, and these indexes must attempt an aggregation of dissimilar items. Indexes always raise serious questions of intuitive (face) validity, as well as issues of measurement reliability and validity. They are usually controversial.

The Arthritis, Rheumatism and Aging Medical Information System (ARAMIS) outcome assessment paradigm has always included a major dimension of treatment toxicity[1]. It was inevitable, therefore, that when we began systematic approaches to pharmacoepidemiology[2] we would need to use a toxicity index, since we sought to compare different drugs, to look at effects over time, to examine effects of patient covariates, and so forth.

However, our initial literature search, in 1989, came up empty. No one had proposed methods of presenting a single index number representing drug toxicity. The Food and Drug Administration could not explain how it determined that a drug was unacceptably toxic, although case study suggested that discovery of a rare serious toxicity might lead to drug recall or that an increase in specific toxicity (e.g., transaminitis) over alternatives might lead to nonapproval. Thus, a drug with greater overall toxicity might sometimes be preferred to one of lesser overall toxicity. Perhaps attempts at overall quantitation represented a fool's errand, an attempt to quantify the unquantifiable. Or, perhaps, no one had used good datasets to approach the issue.

We believe strongly that the issues underlying rational clinical choice are too important to leave fallow simply because they are difficult or because they admit only to imperfect solutions. We have large, prospective, longitudinal datasets including patient reported toxicity, physician recorded toxicity, laboratory values, and with examination of all hospitalizations, and all deaths.

For the past several years, therefore, the ARAMIS Post-Marketing Surveillance Program has been developing methodology to approach a central question of therapeutics: Which drugs are most toxic, and by how much? We have developed and reported on a Toxicity Index that counts and weights symptoms, laboratory side effects, and hospitalizations both by side effect type and severity, and computes these into a single index number representing the toxicity of a particular medicine; this number is adjusted statistically for differences in length of time on different drugs and for differences in characteristics of patients receiving different drugs[3].

MATERIALS AND METHODS

Initial studies have consisted of analysis of many thousands of courses of disease modifying antirheumatic drugs (DMARD), nonsteroidal antiinflammatory drugs (NSAID), and prednisone therapy in 2747 patients with rheumatoid arthritis (RA) over about 8000 patient years of observation. Patients are studied prospectively and longitudinally. All patients have RA, and 5 databank centers were initially involved. Two centers (Santa Clara County and Saskatoon) represent community based populations, 2 (Wichita and Phoenix) represent private rheumatologic practices, and one center (Stanford) represents a university referral practice. Data consist of routine clinical information, demographics, diagnoses, symptoms, physical signs, laboratory findings, therapy employed, and data each 6 months from the Health Assessment Questionnaire (HAQ) detailing disability, symptoms, side effects, and economic effects. Development, validation, and methodology of the

From the Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305.

Supported by grant AM21393 from the National Institutes of Health to ARAMIS (Arthritis, Rheumatism and Aging Medical Information System).

J.F. Fries, MD, Associate Professor of Medicine, Stanford University School of Medicine.

Address reprint requests to Dr. J. Fries, 1000 Welch Road, Suite 203, Palo Alto, CA 94304-1808.

Toxicity Index have been reported. Hospitalizations are individually abstracted from discharge summaries, and deaths from discharge summaries and death certificates. Deaths in patients lost to followup are identified by use of the National Death Index[4,5].

The Toxicity Index includes components of symptoms, laboratory tests, and hospitalizations. Events resulting in hospitalization or death are attributed to a drug if the attending physician made that attribution. If no such attribution was made, they are fractionally attributed on the basis of the relative risk of the event while using this class of medications compared with the relative risk of the event while not using this class of medications. If more than one potential contributing drug is present, as for example with the upper gastrointestinal (GI) hemorrhage associated with both prednisone and an NSAID, attribution is further fractionally allocated between the potential offending agents. Symptoms are additionally categorized as mild, moderate, or severe, and scores weighted accordingly. Units for the Toxicity Index are scored per year of exposure to the agent.

Hospitalization records and death certificates for all events occurring during each 6 month reporting period are reviewed to determine the relationship of hospitalizations and death to drug use. Discharge summaries are supplemented by ARAMIS clinical data and HAQ data. To be attributed, events have to be either directly attributed by a physician to a specific drug, to a specific drug and a clinical problem, or likely to be drug related based upon a known relationship (previously reported in the literature) between a class of drugs and a particular set of symptoms or complications. Thus, any toxic event, no matter how rare, can be attributed if the physician attending the patient believes the drug to be responsible, while in the absence of physician notation, a probable drug relation can be counted as well, but only if the relationship between drug and event is established. The algorithms may be readily applied by trained abstractors in most instances, and in ambiguous situations records are further reviewed by an ARAMIS physician who makes the final judgment[3].

The Toxicity Index is computed as the sum of symptom side effects, laboratory side effects, and hospital days resulting from adverse reactions. Symptom side effect weights range from 0 to 10 and are multiplied by the severity factor, 0.5 for mild, 1.0 for moderate, and 1.5 for severe. The same procedure is employed for laboratory side effects. The number of hospital days is multiplied by the weighting factor for a hospital day (8.4) and fractionated according to the rules described above.

Since some side effects tend to occur early in the course of treatment (e.g., rash), others late (e.g., osteoporosis), and others relatively evenly over time (GI hemorrhage), toxicity index values and standard errors are computed for 6-month periods of exposure (for the first 6 month period, second and third 6 month periods, etc.) for each drug and then combined. Raw scores are statistically adjusted by standardization as described below, using age, race, sex, duration, disease severity, and new start versus continuing therapy, comorbidity, presence of concurrent therapy, and other variables. Sensitivity analyses using several alternative weighting systems have been performed, and yield similar results.

Regression trees were used to develop strata for standardization, following the procedures of Bloch and Segal and using the computer program CART[6]. Regression trees were constructed for the first period, the second and third periods, periods 4 through 8, and periods 9 and above. A similar list of classifying variables was obtained for each set of periods. Each regression tree first split either on disability or comorbidity, and these variables appeared in all trees. Additional variables selected included age, duration, and number of swollen joints. Since we are not able to distinguish patients who were first prescribed a drug in the first observation period from those who had begun taking the drug before our observation, we repeated the standardized analyses on those who were not taking the drug at first observation period but subsequently began treatment. The variables appearing in these regression trees were closely similar to those already described. Strata developed by these techniques were used to develop standardized toxicity index scores. The hypothesis for analysis is that there is no difference in standardized toxicity index scores between drugs; thus, the alternative hypotheses are 2-sided. The test statistic is based on the normal approximation to the distribution of differences between the standardized toxicity index values.

RESULTS

Table 1 presents the toxicity indices for NSAID[4,7]. There is a range of 3 to 4 fold in toxicity between the most toxic and least toxic NSAID. Of prescription NSAID, when employed for RA, salsalate, ibuprofen, and naproxen are the least toxic, and ketoprofen, tolmetin, meclofenamate, and indomethacin are the most toxic.

Table 2 shows a similar display of relative toxicity for the DMARD[5,7]. A wide range of toxicities is seen here, ranging from a benign profile and score for hydroxychloroquine to higher values for methotrexate, azathioprine, auranofin, and the reference drug prednisone. Comparing data for NSAID (Table 1) with DMARD (Table 2), interesting conclusions emerge. All data were obtained from the same patients over the same time period, with statistical adjustment for time taking the drug and for characteristics of patients receiving specific therapies. The overlap of toxicities between NSAID and DMARD is much more pronounced than are the differences, both before and after statistical adjustment. Hydroxychloroquine would be a very nontoxic NSAID, while the most toxic NSAID exhibit similar overall toxicity as DMARD such as methotrexate and azathioprine. Each of the results reported above holds when the overall drug experience or only new starts are analyzed, and holds at each of 5 clinical centers with greatly differing

Table 1. *Relative toxicity of NSAID: data from 5 ARAMIS databank centers; standardized toxicity index scores*

Drug	Number of Courses	Mean ± SE	Rank
Salsalate	121	1.28 ± 0.34	1
Ibuprofen	503	1.94 ± 0.43	2
Naproxen	939	2.17 ± 0.23	3
Sulindac	511	2.24 ± 0.39	4
Piroxicam	790	2.52 ± 0.23	5
Fenoprofen	161	2.95 ± 0.77	6
Ketoprofen	190	3.45 ± 1.07	7
Meclofenamate	157	3.86 ± 0.66	8
Tolmetin	215	3.96 ± 0.74	9
Indomethacin	386	3.99 ± 0.58	10

Table 2. *Relative toxicity of DMARD: data from 5 ARAMIS databank centers; standardized toxicity index scores*

Drug	Number of Courses	Mean ± SE	Rank	Hospitalization Component
OH-Chloroquine	639	1.38 ± 0.15	1	0.00
Intramuscular gold	659	2.27 ± 0.17	2	0.15
D-penicillamine	496	3.38 ± 0.36	3	0.99
Methotrexate	660	3.82 ± 0.35	4	1.02
Azathioprine	190	3.92 ± 0.39	5	0.83
Auranofin	409	5.25 ± 0.32	6	0.00

patient selection characteristics. Rank orders also persist after extreme changes in weighting are employed, from totally unweighted to using squared weights.

In the comparative DMARD experience reported above, 2 drugs, hydroxychloroquine and auranofin, were not associated with serious toxicity requiring hospitalization or resulting in death, although auranofin had a very frequent occurrence of annoying but reversible side effects, principally diarrhea. Additionally, when a toxicity index for NSAID is constructed from only GI toxicity, rank orders are essentially identical.

DISCUSSION

Toxicity index information is important and provides unique insights, in part because of the methodologic difficulties. A recent Newcastle conference on NSAID toxicity brought together 12 groups with data on comparative toxicity (Henry, unpublished data). Despite differences in datasets, drugs studied, methods, and assumptions, the data were congruent. For example, all groups had studied ibuprofen, naproxen, and piroxicam, and in every study ibuprofen was least toxic, naproxen middle, and piroxicam most toxic. Moreover, the range of toxicity, about 3 fold over all drugs, was similar. Thus, data appear valid.

On the other hand, aspirin was unexpectedly nontoxic in our data, leading to further study. The relative benignity of aspirin was found to be accounted for by a low relative dose (2665 mgm/day) and by frequent use of coated aspirin preparations in actual clinical experience[8]. Thus, actual experience may be different from that predicted in preapproval clinical trials.

Remaining caveats and requirements are numerous. To avoid hidden effects of underlying assumptions, raw data must be presented when data are published so that others may recompute with different assumptions. Better weighting systems are desirable; although they are unlikely to change results very much they may better satisfy purists. We have now changed our statistical adjustment procedures for differing patient characteristics when taking different drugs, using general linear models for analysis of covariance; while numbers change, actual results are closely similar.

The standard procedure calculates the crude toxicity index for each drug as the total toxicity units accrued for that drug divided by the total years of exposure to the drug, then statistically adjusts the crude index for patient characteristics including age, sex, time on drug, previous side effects, disease duration, educational level, disability level, concurrent medications, and others, using analysis of covariance with SAS software (SAS Inc, Cary, NC). The new procedure calculates a toxicity index score for each patient consisting of the toxicity units for a drug divided by the length of time taking drug, averages these scores across patients, then statistically adjusts scores similarly for differing patient characteristics.

Nevertheless, most toxicity studies will be observational if data are to be sufficient, and if patients receiving different drugs are quite different, statistical adjustment is likely to be incomplete. This does not appear a problem with NSAID, but may be with DMARD. Death is the ultimate side effect, but occurs too seldom to be included in an index even if the weighting problems could be solved. Compliance and relative dosage present additional problems, and dose/side effect relationships are needed for all drugs studied.

Yet the rewards are many. The stage is set for quantitative study of toxic–therapeutic ratios (e.g., toxicity index/disability index changes) that can further inform clinical choice. From a health policy perspective, increased use of less toxic NSAID and increased avoidance of the most toxic can importantly reduce aggregate NSAID toxicity, perhaps by one-half or more[4,7,9–11].

REFERENCES

1. Fries JF: Toward an understanding of patient outcome measurement. *Arthritis Rheum 1983;*26:697–704.
2. Fries JF, Singh G, Lenert L, Furst DE: Aspirin, hydroxychloroquine, and hepatic enzyme abnormalities with methotrexate in rheumatoid arthritis. *Arthritis Rheum 1990;*33:1611–9.
3. Fries JF, Spitz PW, Williams CA, Bloch DA, Singh G, Hubert HB: A toxicity index for comparison of side effects among different drugs. *Arthritis Rheum 1990;*33:121–30.
4. Fries JF, Williams CA, Bloch DA: The relative toxicity of nonsteroidal antiinflammatory drugs. *Arthritis Rheum 1991;*34:1353–60.
5. Fries JF, Williams CA, Bloch DA: The relative toxicity of disease modifying antiinflammatory drugs. *Arthritis Rheum 1991;*34:1353–60.
6. Breiman L, Freidman J, Olshen R, Stone C: *Classification and Regression Trees.* Belmont, CA: Wasworth, 1984:93–129.
7. Fries JF, Williams CA, Ramey DR, Bloch DA: The relative toxicity of alternative therapies for rheumatoid arthritis: Implications for the therapeutic progression. *Semin Arthritis Rheum 1993;*23:68–73.
8. Fries JF, Ramey DR, Singh G, Morfeld D, Bloch DA, Raynauld JP: A reevaluation of aspirin therapy in rheumatoid arthritis (RA). *Arch Intern Med 1993;*153:2465–71.
9. Gabriel SE, Bombardier C: NSAID induced ulcers: An emerging epidemic? *J Rheumatol 1990;*17:1–4.
10. Fries JF, Miller SR, Spitz PW, Williams CA, Hubert HB, Bloch DA: Toward an epidemiology of gastropathy associated with nonsteroidal antiinflammatory drug use. *Gastroenterology 1989;*96:647–55.
11. Fries JF, Williams CA, Bloch DA, Michel BA: Nonsteroidal antiinflammatory drug-associated gastropathy: Incidence and risk factor models. *Am J Med 1991;*91:213–22.