Critical Issues in Longitudinal and Observational Studies: Purpose, Short Versus Long Term, Selection of Study Instruments, Methods, Outcomes, and Biases

FREDERICK WOLFE

ABSTRACT. Longitudinal observational studies (LOS) provide key information about outcomes and treatment effectiveness that are not available from other types of investigations, including randomized controlled trials. Although LOS are easy to perform, they are difficult to perform correctly. Major problems include recruitment, retention, and relevance, but the central problems in LOS are bias, understanding the nature of the biases, and reporting the biases. To advance the quality and validity of LOS there must be a uniform requirement for reporting detailed data that includes details about the sampled population and the sampling methods, the rationale for the selection of control subjects, the probable biases, and estimates of the extent and consequences of the biases. Sufficient covariates should be collected so that where possible statistical adjustment for bias can be made. (J Rheumatol 1999;26:469–72)

Key Indexing Terms:
LONGITUDINAL OBSERVATIONAL STUDIES OUTCOMES
RHEUMATOID ARTHRITIS OSTEOARTHRITIS STUDY REQUIREMENTS

Science does not begin with a tidy question, nor does it end with a tidy answer. — John Tukey

Previous OMERACT recommendations for rheumatoid arthritis and osteoarthritis were designed for use in randomized clinical trials (RCT) of short or intermediate durations to assess the comparative efficacy of different treatments. Observational studies and longitudinal observational studies (LOS) have substantially different goals, assessment methods, potential biases, and problems, as shown in Table 1.

Studies and reports. LOS generally refer to investigations designed to answer a series of related questions or test a series of related hypotheses. Although it is possible that a study will produce a single manuscript or report, more often the study project will produce a number of allied reports. For example, the Framingham study forms the basis of a number of specific reports or sub-studies[1]. Implicit in this broad definition is the multipurpose nature of LOS and, consequentially, the multi-fold variable collection process. For example, an investigation may collect mortality data at the same time it is collecting data relating to drug effectiveness.

Duration. Because of the duration of LOS, investigators may expect substantial dropouts due to non-compliance, other illnesses, and death. Unlike RCT, these dropouts are non-random, and tend to affect the older and younger, men more than women, those with more severe rheumatic disease, the anxious or depressed, and the less well educated. Because of dropout and missing data problems and the longitudinal nature of the data, LOS may require complex statistical methods[2] that may include the use of time varying covariates. This in turn implies that appropriate covariates must be collected. Although subject loss is inevitable, it can be reduced by rigorous followup, and by reduction of respondent burden by simplifying questionnaires and assessments.

Study goals, study types, and study limitations. Unlike RCT, the goals and methods of LOS vary widely. One (sub)-study may address costs, another coping ability, mortality, or radiographic progression; and radiographic scores may not be of interest to investigators studying costs, nor the psychologists studying coping. LOS may be conducted in the clinic where the collection of detailed questionnaire assessments may be difficult. Questionnaire surveys, on the other hand, may not have access to serial physical examination or laboratory data. Similarly, data bank investigators may be interested in all variables that describe or influence outcome, but must reduce their purview because of limited funding. External factors, then, as well as investigator interest define the nature of LOS.

Observational studies lasting more than 2 years will have a different set of outcomes of interest. Work disability or mortality, for example, are important outcomes for longer but not for shorter studies. Therefore a single "core set" of variables cannot be expected to include all important LOS variables, and a secondary set of "potentially important items" is required. Core variables will differ according to the length of the study and will be those variables that, given the nature of the study, should almost always be collected. Secondary variables are those items that are either less

From the University of Kansas School of Medicine, Wichita, Kansas, USA.

F. Wolfe, MD, Clinical Professor of Internal Medicine, Arthritis Research Center and University of Kansas School of Medicine.

Address reprint requests to Dr. F. Wolfe, Arthritis Research Center, 1035 N. Emporia, Suite 230, Wichita, KS 67214; E-mail: fwolfe@southwind.net

Table 1. Comparison of randomized controlled trials and longitudinal observational studies.

Item	Longitudinal Study	RCT
Duration, yrs	2–25	3 mo–2 yrs
Major assessment	Disease activity & non-disease activity outcomes 1. Pain 2. Functional disability 3. Fatigue and sleep status 4. Radiographic progression 5. Damage 6. Work disability 7. Psychological status 8. Economic status 9. Mortality	Disease activity 1. Pain 2. Function 3. Joint examination 4. Radiographic progression in 2 year studies
Examination method	Multiple methods 1. Clinical examination 2. Regular or periodic questionnaire assessments 3. Surveys	Single method 1. Clinical examination
Funding	May be poor or good	Usually very good
Feasibility of assessments	Often important	Usually not an issue
Covariates & baseline variables	Important	Removed by randomization process
Subject loss	Often substantial: by mortality and subject refusal	Usually not an issue
Biases	Potentially many, including selection	Spectrum
Statistical analyses	Often complex	Generally simple

important to LOS or whose inclusion depends upon special study interests.

Practicality/feasibility. The most detailed and/or the most accurate measure is often not the best measure. Tukey observed that "simplicity and flexibility outweigh efficiency...," and that the "... 'practical power' of a statistical test... [is] the product of the probability that the test will be applied and the mathematical power when applied"[3]. During patient assessment, either in the clinic or by survey, the investigator has only just so much time, after which patients become resistant or refuse assessments, and are more likely to decline future study. Therefore, shorter questionnaires in some areas may allow for broader and more useful assessment overall. In the clinic one cannot easily repeatedly administer instruments such as the Sickness Impact profile or the Arthritis Impact Measurement Scale 2[5]. Wolfe and Pincus and others have collected longitudinal data in the clinic during routine clinic care using short questionnaires and assessments where longer and more detailed assessments would have been impractical[6,7]. These concerns extend to examination as well. Detailed joint assessments of swelling and tenderness as well as measurements of range of motion or strength are expensive and time consuming.

Unlike the RCT where a few common tests are utilized, no one test or method is appropriate in all situations. Investigators may wish to use shorter or longer, less or more detailed methods, according to the study purpose and design.

Relevance. To a large extent RCT address issues of process: changes in joint swelling, joint tenderness, pain, function, global assessments and, for RA, acute phase reactants. LOS studies more often address issues of outcome: status in regard to functional and work disability, costs, socioeconomics, or mortality. Measurements that may be useful in RCT, such as pain, global severity, and joint counts, appear to be considerably less useful as longitudinal measure than function, costs, service utilization, and mortality[8,9].

Which instrument? Which score? Depending on the rheumatic condition under study, a number of different questionnaires are available to measure similar concepts. The comparative performance of each instrument in observational and longitudinal studies should be the subject of a future conference regarding instrumentation. Issues of importance include not only validity and reliability in the LOS setting, but also a review of the psychometric properties of the instruments. Tennant, *et al*[10] and Stucki, *et al*[11] have presented data that instruments such as the Health Assessment Questionnaire (HAQ)[12] represent ordinal scales

that may require new scaling and analysis techniques, and that construct validity may be impaired when instruments such as the HAQ are used in disorders other than RA[10]. These concerns are likely to be true with other questionnaires commonly used in rheumatic diseases. While one instrument might be better than another, issues of feasibility, national and language differences, and differences in study purposes suggest that many instruments are acceptable.

Problems in RCT and LOS. RCT are designed to detect differences between interventions. The usual rheumatic disease outcome to be detected is a difference in clinical status. In RA, clinic status means a change in disease activity; in OA clinical status means pain and function; in syndromes such as fibromyalgia it might be defined to mean pain, fatigue, and psychological distress. In longer duration clinical trials changes in radiographic progression might also be measured. RCT become increasingly less practical as the duration of study is increased, primarily for 3 reasons: increased dropouts, high costs, and questions of relevance. LOS that are not very good at answering the short term questions asked by RCT become increasingly relevant at asking these and other questions with increasing followup duration. But LOS cannot usually be used to detect differences between interventions because of the non-random assignment of the interventions.

RCT are inherently biased. That is one of their initial strengths — and longer term weaknesses. For example, to maximize statistical power and minimize sample size and costs, patients are selected for RCT on the basis of having high levels of disease activity (or pain and dysfunction in OA). In general, study patients tend to be at the 60th–80th percentile of disease activity[13]. This makes sense. If you want to study inhibition of radiographic progression or substantial decrease in disease activity, then you must have subjects who are capable of easily displaying that change. But there is a price to pay for selecting such subjects, and that is that the results are usually not generalizable to the majority of rheumatic disease patients. In addition, RCT screen out really sick patients such as those with renal, cardiac, or gastrointestinal disease — patients who will be treated in the clinic, however, regardless of their medical status.

The RCT randomization process "homogenizes" patients. The RCT doesn't care if education level, age, sex, income level, or even genetic markers convey benefit to patients, since the nature of the randomization process is to distribute these factors randomly and usually equally among the study arms.

Table 2. Biases in longitudinal observational studies (LOS).

Bias	Example	Consequence	Required Reporting	Potential Problems
Center (spectrum) bias	Disease severity, age, economic status, insurance status (US), referral characteristics	Differing prognosis and outcome	Describe patient population and give comparative data	If you tell reviewers your problems you won't get your paper accepted
Recruitment or selection bias (therefore spectrum bias)	As above, and difference in disease severity	Differing prognosis and outcome	Full description of recruitment procedures and give comparative data	If you tell reviewers your problems you won't get your paper accepted
Control bias	Inappropriate controls and selection bias in obtaining controls	Meaningless results	Full discussion of methodology and appropriateness of controls	If you tell reviewers your problems you won't get your paper accepted
Loss to followup	As above, difference in socio-demographic characteristics and disease severity	Differing prognosis and outcome	Full description of subjects, and description of statistical methods to help to control for loss	If you tell reviewers your problems you won't get your paper accepted
Questionnaire vs clinic bias	Different groups of subjects and differences in socio-demographics and severity	Results differ according to the setting in which the data are collected	Full description of comparative samples and discussion of consequences of using current sample	If you tell reviewers your problems you won't get your paper accepted
Inadequate followup	Short term study of mortality or disability	Results may be wrong	Description of rationale of study termination	If you tell reviewers your problems you won't get your paper accepted
Assessment bias	Lack of blinding, non standardized and nonvalidated assessment methods	Uninterpretable results	Full description of study methods	If you tell reviewers your problems you won't get your paper accepted

The central problems in LOS are bias, understanding the nature of the biases, and reporting the biases (Table 2). One frequently comes across a biased study in which the "potential" bias is mentioned in the discussion section of the manuscript, as if a mention of the problem is sufficient to repair it. An additional problem regarding the report is to generalize from a biased sample to the universe of patients with a rheumatic disease or, to paraphrase Schopenhauer, taking the limits of our own field of vision for the limits of the world.

Examples of center bias are rheumatology centers where the sex distribution of patients may be markedly imbalanced (e.g., US veterans' hospitals), clinics that serve the poor, the rich, or that have restrictions on entry based on some economic criterion.

Perhaps the most serious bias is selection or recruitment bias. Examples include the use of convenience samples, samples based on cases attending clinic versus potential cases not currently attending clinic. This example of prevalence versus incidence bias leads to the identification of patients with more serious medical or psychological problems. Samples of patients completing surveys are systematically different from persons not completing surveys[14,15], and survey participants differ from patients attending clinics[16]. The use of patients who are referred for a special purpose almost always results in seriously biased samples. Even within clinics, the failure to use consecutive cases or a method of random identification of cases also can lead to seriously biased samples. Selection bias is often the major explanatory factor for the study result. In case–control studies and other studies employing controls, the same issues of selection bias are important, in addition to that of the appropriateness of the controls. Loss to followup in the clinic population and dropouts in LOS can lead to a systematic selection bias. Some biases are unavoidable, for example, spectrum bias in a clinic devoted to serving the poor. But sampling biases are largely avoidable by better study design.

What to do about the problem? Many papers in the literature are uninterpretable or of limited usefulness, or even wrong, because of problems of bias. Yet they may be cited often or used for evidence for or against a hypothesis. The current publication and review process has not been able to prevent publication of inadequate studies, for there is no uniform requirement for reporting on study design and methodology; and there is yet no agreement as to what constitutes an acceptable LOS. Until such requirements are developed authors will continue to conceal study defects rather than avoid them.

What to do. To advance the quality and validity of LOS there must be a uniform requirement for reporting detailed data that includes details about the sampled population and the sampling methods, the rationale for the selection of cor trol subjects, the probable biases, and estimates of the exte and consequences of the biases. Sufficient covariates shoul be collected so that where possible statistical adjustment fc bias can be made. Insisting on these requirements woul improve the quality of studies because authors would begi to observe appropriate methods in order to have their pape published.

REFERENCES

1. Felson DT, Zhang YQ, Hannan MT, et al. Risk factors for incident radiographic knee osteoarthritis in the elderly: The Framingham Study. Arthritis Rheum 1997;40:728–33.
2. Diggle PJ, Liang KL, Zeger SL. Analysis of longitudinal data. 1st ed. Oxford: Clarendon Press; 1994.
3. Exploratory data analysis as part of a larger whole. In: Tukey JW, Jones LV, editors. The collected works of John W. Tukey. Monterey Wadsworth & Brook; 1984:799.
4. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787–805.
5. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis Rheum 1992;35:1–10.
6. Wolfe F, Pincus T. Data collection in the clinic. Rheum Dis Clin North Am 1995;21:321–58.
7. Pincus T, Wolfe F. Patient self-report questionnaires as integral to clinic care. Behav Meas Lett 1997;53–7.
8. Hawley DJ, Wolfe F. Sensitivity to change of the Health Assessme Questionnaire and other clinical and health status measures in rheumatoid arthritis: results of short term clinical trials and observational studies versus long term observational studies. Arthritis Care Res 1992;5:130–6.
9. Wolfe F, Hawley DJ, Cathey MA. Clinical and health status measures over time — prognosis and outcome assessment in rheumatoid arthritis. J Rheumatol 1991;18:1290–7.
10. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol 1996;35:574–8.
11. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol 1996;49:711–7.
12. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
13. Wolfe F. The prognosis of rheumatoid arthritis: assessment of disease activity and disease severity in the clinic. Am J Med 1997;10312–8.
14. Allebeck P, Ahlbom A, Allander E. Increased mortality among persons with rheumatoid arthritis, but where RA does not appear o death certificate: eleven year follow-up of an epidemiological stud Scand J Rheumatol 1981;10:301–6.
15. Edworthy SM, Martin L, Barr S, Birdsell DC, Brant RF, Fritzler M A clinical study of the relationship between silicone breast implant and connective tissue diseases. J Rheumatol 1998;25:254–60.
16. Hawley DJ, Wolfe F. Effect of light and season on pain and depression in subjects with rheumatic disorders. Pain 1994;59:227–34.