Methodological Issues in Conducting and Analyzing Longitudinal Observational Studies in Rheumatoid Arthritis

DEBORAH P.M. SYMMONS

ABSTRACT. This article discusses 5 methodological issues that arise in the course of conducting longitudinal observational studies: generalizability, missing data, repeated measures on the same individual, measures taken at varying time points from symptom onset, and assessing the effect of treatment. Methods discussed include general estimating equations and propensity scores. The points are illustrated by examples from the Norfolk Arthritis Register dataset. (J Rheumatol 2004;31 Suppl 69:30–34)

Key Indexing Terms:
RHEUMATOID ARTHRITIS
REPEATED MEASURES
OBSERVATIONAL STUDIES
PROPENSITY SCORE

In longitudinal studies a group of individuals is followed for a period of time and repeated measurements are performed on those individuals. Longitudinal studies give rise to longitudinal data (also known as repeated measures data). Randomized controlled trials (RCT) are longitudinal studies. Longitudinal observational studies (LOS) are those in which patient management is decided on clinical grounds rather than being assigned by randomization. LOS have a number of advantages over RCT. First, they are more generalized because there are no exclusions with regard to enrolment. Second, it is possible to maintain followup over long periods of time, whereas it is difficult to maintain blinding and treatment assignment in an RCT for more than 3 years.

This article examines 5 methodological issues in conducting LOS involving patients with rheumatoid arthritis (RA) (Table 1). Problems and potential issues are illustrated using data from the Norfolk Arthritis Register (NOAR). NOAR is a primary-care based inception cohort of patients with recent onset inflammatory polyarthritis (of which RA is a major subset)[1]. NOAR was started in 1990 and patients registered in that year have now been followed for over 10 years. There are currently over 3500 individuals on the Register.

Table 1. Methodological issues in the analysis of longitudinal observation studies.

- Generalizability
- Loss to followup and missing data
- Repeated measures on each individual
- Observations at different time points relative to symptom onset
- Non-random assignment to treatment

GENERALIZABILITY

Most investigators conducting LOS in RA hope that the results will apply, not only to their own patients, but to others in their own country. However, as with all epidemiological studies, the results of a LOS can only be generalized to other patients who would have satisfied the original entry requirements of the LOS and who have similar characteristics to those patients who remained under followup (and could therefore be included in the analysis). It is therefore important that all published LOS should specify their catchment population (usually a particular geographical area), the source of patients (e.g., primary or secondary care), entry criteria (e.g., satisfying the 1987 ACR criteria[2]), and loss to followup (Table 2). The LOS investigators should endeavor to enrol all patients they see who satisfy the entry criteria and should specify how many patients they forgot to ask or who declined to participate. The issue of reporting requirements for LOS in rheumatology was considered at the 4th Outcome Measures in Rheumatology Clinical Trials (OMERACT) meeting in 1998[3].

LOSS TO FOLLOWUP AND MISSING DATA

Many statistical modelling tests can only include individuals with complete sets of data. Yet all LOS will have some loss to followup and missing data. If 100 patients with RA were followed for 10 years and each patient completed a Stanford Health Assessment Questionnaire (HAQ) (which comprises 20 questions)[4] each year, the study should have amassed 20,000 items of data. If only 1% of items (200) are missing, there may be no patients with complete sets of data.

From the ARC Epidemiology Unit, School of Epidemiology and Health Sciences, The University of Manchester, Manchester, United Kingdom.

Funded by the Arthritis Research Campaign, UK.

D.P.M. Symmons, MD, MFPH, FRCP, Professor of Rheumatology and Musculoskeletal Epidemiology.

Address reprint requests to Dr. D.P.M. Symmons, ARC Epidemiology Unit, School of Epidemiology and Health Sciences, Stopford Building, The University of Manchester, Oxford Road, Manchester M13 9PT, UK. E-mail: deborah.Symmons@man.ac.uk

Table 2. Recommended reporting requirements for longitudinal observational studies in RA. From Wolfe, *et al*. J Rheumatol 1999; 26:484–9.

Item	Information
Study design	Prospective, retrospective, or mixed
Source of cases	True population based, catchment population, or consecutive series
Timing of recruitment	In relation to disease onset, first presentation, or prevalent cases
Inclusion criteria	Examples: classification criteria for RA, age, ethnic group
Demographic data collected	Examples: age, gender, ethnic group, socioeconomic group
Baseline clinical data collected	Whether collected prospectively or from medical record review.
	Items collected. Interobserver variation.
	Numbers of individuals with missing data
Followup data collected	Frequency of followup items collected.
	Numbers of individuals with missing data.
	Method of followup data collection (e.g., record review, telephone contact)
Analyses	Methods and rationale.
	Power to detect clinically meaningful change.
	Handling of missing data.
	Tests of internal and external validity

It is therefore very important, first to maximize the completeness of data collection, and second to understand and adjust for missing data. Data are said to be missing completely at random (MCAR) if there are no discernible differences between those individuals who have a particular item of data and those who do not. In this situation the individuals missing that item of data are a random subset of the whole cohort. This is seldom the case. For example, patients usually withdraw from a study for reasons related to their disease status. Data are said to be missing at random (MAR) if they are missing conditional on an observed value. For example, men may be less likely than women to attend for radiographs. However, within the strata (men and women) there is no discernible difference (based on observed or unobserved values) between those who did or did not attend for a radiograph. Radiographic data would then be said to be MAR. This situation can be compensated for by weighting. For example, if only half the men attend for a radiograph, then the results from each man who does attend can be weighted by 2.

Data that are MCAR or MAR are said to be "ignorable," although ignoring them may severely reduce the power of the study. Data not missing at random, that is, where the missing values are dependent on variables that have not been measured (i.e., it is impossible to get down to a stratum within the dataset in which the data are then MAR) are said to be non-ignorable.

There are various ways of imputing data that are MCAR or MAR. This helps to restore the power of the study. Occasionally the missing data can be correctly inferred from the remaining data. For example individuals aged under 16 may be assumed to be unmarried. Otherwise there are 2 main ways of imputing data. The first is called "deterministic" and is based on using actual values, e.g., last observation carried forward (often used in RCT), mean of non-missing values for that value in the same stratum, or regression to predict the missing value. Such techniques, however, lead to overestimation of the precision of the final estimate (the 95% confidence interval will be artifically narrow). The second group of techniques, stochastic imputation, retains an element of random variation. Examples include hot decking and regression with simulated error. Imputed values should always be flagged within a data set.

Various procedures have been developed for handling non-ignorable missing data but they are all dependent on strong model assumptions.

Four hundred thirty-three patients were recruited by NOAR in 1990 and 1991. After 5 years, 44 (10%) had died, 47 (11%) had withdrawn from the study, and 24 (6%) had been lost to followup (Table 3). There were systematic differences between those who completed 5 years' followup and those who died (e.g., those who died were older at symptom onset and more likely to be male). Those who withdrew or were lost to followup were also more likely to be male and had milder disease than those who remained in the study (Table 4). This means that, for example, HAQ data over 5 year followup in the NOAR cohort are not MCAR. However, among those who completed 5 year followup there were no discernible differences between those who had a complete HAQ for each of the 5 years and those with one or more missing HAQ scores. These data are therefore MAR.

REPEATED MEASURES

Repeated measures on individuals are not independent. They are correlated. The strongest predictor of the HAQ score in an RA patient one year after presentation is her HAQ score at baseline[5]. Any analysis that does not take into account this lack of independence will be biased, and the standard errors and confidence intervals will be artificially narrow[6].

There are a number of ways of dealing with this issue.

Table 3. Norfolk Arthritis Register — followup of patients registered in 1990–1991.

Time from Registration (yrs)	Patients Assessed (% of Baseline)	Cumulative loss (%) Died	Declined	Lost to Followup
0	433			
1	409 (94)	7	8	2
2	380 (88)	19	17	11
3	362 (84)	25	22	16
4	344 (79)	37	26	19
5	318 (73)	44 (10)	47 (11)	24 (6)

Table 4. Differences between Norfolk Arthritis Register patients followed for 5 years and those who died or were lost to followup.

	Followed, n = 318	Died, n = 44	Lost/Declined, n = 71
Median age at onset, yrs (IQR)	55 (41–55)	70 (64–79)	53 (39–68)
Female, %	67	46	47
Median HAQ (IQR)	0.75 (0.25–1.25)	1.375 (0.375–2.06)	0.75 (0–1.365)
RF positive, %	35	59	31
ACR criteria positive, %	50	46	39

One is to look only at paired data. This was often done in older RCT, where the first and last values in the trial were compared and all measurements in between were disregarded.

Alternatively, the repeated measures may be reduced to a single summary[7]. Examples would be the graphically derived area under the curve (AUC), or the time taken to certain outcome (e.g., first erosion), or the rate of change of a variable over time (e.g., radiological progression). However, these methods may still not fully exploit the available data.

General estimating equations (GEE) are a multivariate extension of generalized linear models[8] that allow for within-subject correlation. They allow the examination of the relationship between multiple prognostic factors and an outcome over time. The odds ratios obtained are presumed to hold constant over time. This assumption can be tested by introducing interaction terms between time and each predictor variable in the GEE model.

Figure 1 shows the association between age at symptom onset (divided into tertiles) and gender and having a HAQ score ≥ 1.0 over 5 years of followup in 684 patients on the NOAR[9]. These are the results of multivariate analysis; the model also included time from symptom onset to presentation, duration of morning stiffness, rheumatoid factor, and number of deformed joints and rheumatoid nodules.

The GEE model uses all available data and assumes that data are MCAR. As noted above, the NOAR HAQ data are MAR. However, a weighted analysis that allowed in particular for the difference in age between those who did or did not complete 5 years' followup yielded very similar results.

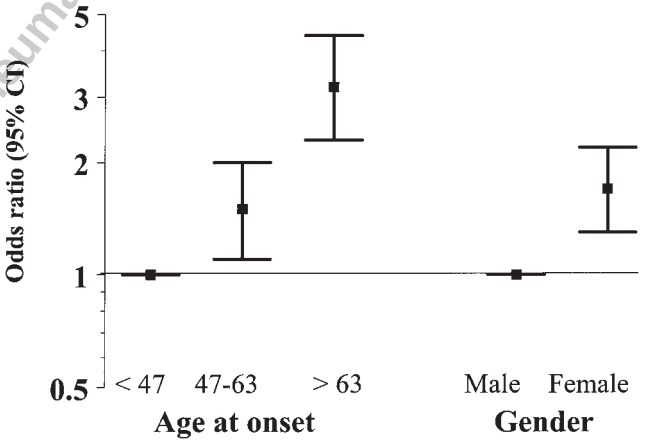

Figure 1. Odds of moderate disability (HAQ ≥ 1.0) at 5 years in the Norfolk Arthritis Register.

OBSERVATIONS AT VARYING TIME POINTS SINCE SYMPTOM ONSET

Most LOS in early RA recruit patients with a fairly wide disease duration (e.g., < 2 years, < 3 years). A set of observations and investigations is recorded at the baseline visit, which is regarded as "time zero." This means that, for example, in a study that has recruited patients with up to 3 years' symptoms, radiographs taken 2 years after baseline may actually be at anything from 2 to 5 years from symptom onset. If it is important to determine when an outcome occurred in relation to disease duration, then the timing of the measurement should be expressed relative to the time of symptom onset. In NOAR this technique was used when

investigating the time of first erosions[10]. Patients were included in this study if they had paired radiographs in which the first set was erosion-free. If erosions were seen in the second set of radiographs, they were assumed to have occurred at time points following a Poisson distribution between the 2 films. All times were measured from recalled symptom onset.

ASSESSING THE EFFECT OF TREATMENT

It is difficult to study the effect of treatment on outcome in LOS because the decision to treat a patient and which drug to use is not random. There are, therefore, likely to be systematic differences in the demographic and disease characteristics of patients who are, and are not, treated. Patients who are treated tend to have more serious disease and a worse prognosis than those judged not to require therapy. Unless the chosen treatment is totally effective in abolishing all evidence of disease activity and cumulative damage, patients who are treated are likely to have a worse outcome than those who are not. This could lead to the conclusion that treatment is harmful, but the worse results in treated patients are actually explained by "confounding by indication." This bias in treatment assignment makes it difficult to assess the therapeutic effect of treatment in LOS.

In recent years a number of statistical techniques have been proposed that adjust for the variables that influence the decision to treat. One of these is the propensity score[11]. The propensity score provides an estimate of the probability that a patient will receive treatment based on the disease characteristics of that patient (for example, erythrocyte sedimentation rate, tender and swollen joint count, rheumatoid factor status). Providing that the propensity score is based on a large number of the measures that influence the decision to treat, and providing the prediction is accurate, then the estimate of the effect of treatment after adjusting for the propensity score can be viewed as unbiased. Within a group of patients with the same propensity score, treatment assignment can be viewed as random. In other words, among patients with a propensity score of 0.9 (i.e., a 90% probability of treatment) it is random whether an individual is in the 90% who are treated, or the 10% who are not. So adjusting for the propensity score is effectively adjusting for pretreatment disease severity and is the equivalent of an intention-to-treat analysis in a randomized controlled clinical trial.

Using propensity scores, it has been possible in the NOAR cohort to show, for the first time in a LOS, that early treatment with disease modifying drugs (DMARD) or steroids reduces the probability of being significantly disabled at 5 years to below that in the group of patients judged not to require such treatment[12] (Figure 2). The same approach has been used to show that early DMARD therapy is beneficial with regard to radiographic outcome[13].

CONCLUSION

In conclusion, longitudinal observational studies in RA can provide valuable information on outcome. They tend to be more generalizable and can follow patients for longer than randomized clinical trials. They present interesting analytic challenges, in particular with regard to missing data, repeated measures, and assessing the effect of treatment. Many of these challenges are now solvable with large datasets and modern computerized statistical methods. The problems of LOS are not unique to rheumatology and we should learn from and collaborate with those working with patients with other chronic diseases.

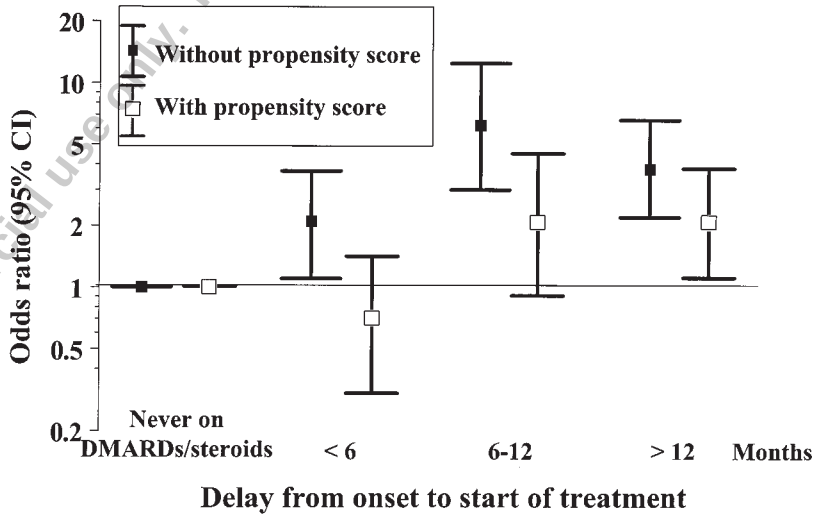

Figure 2. Association between demographic variables and HAQ ≥ 1.0 at 5 years in the Norfolk Arthritis Register.

ACKNOWLEDGMENT

I would like to thank Alan Silman, my co-principal investigator and Marwan Bukhari, Graham Dunn, Beverley Harrison, and Nicola Wiles for valuable discussions about these methodological issues and their potential solutions. We are grateful to the general practitioners and rheumatologists in the Norwich area for continuing to refer patients to and support NOAR.

REFERENCES

1. Symmons DPM, Barrett EM, Bankhead CR, Scott DGI, Silman AJ. The incidence of rheumatoid arthritis in the United Kingdom: results from the Norfolk Arthritis Register. Br J Rheumatol 1994;33:735-9.
2. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
3. Silman AJ, Symmons DPM. Reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:481-3.
4. Fries JF, Spitz P, Kraines RG, Holman HR, Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
5. Harrison BJ, Symmons DPM, Brennan P, et al. Inflammatory polyarthritis in the community is not a benign disease: predicting functional disability one year after presentation. J Rheumatol 1996;23:1326-31.
6. Ward MM, Leigh JP. Pooled time series regression analysis in longitudinal studies. J Epidemiol 1993;46:645-58.
7. Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ 1990;300:230-5.
8. Bull SB. Regression models for multiple outcomes in large epidemiologic studies. Stat Med 1998;17:2179-97.
9. Wiles N, Dunn G, Barrett E, Silman A, Symmons D. Association between demographic and disease related variables and disability over the first five years of inflammatory polyarthritis: a longitudinal analysis using generalised estimating equations. J Clin Epidemiol 200;53:988-96.
10. Bukhari M, Harrison B, Lunt M, Scott DGI, Symmons DPM. Time to first occurrence of erosions in inflammatory polyarthritis. Arthritis Rheum 2001;44:1248-53.
11. D'Agostino RBJ. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265-81.
12. Wiles NJ, Lunt M, Barrett EM, et al. Reduced disability at five years with early treatment of inflammatory polyarthritis: results from a large observational cohort using propensity models to adjust for disease severity. Arthritis Rheum 2001;44:1033-42.
13. Bukhari MAS, Wiles NJ, Lunt M, et al. Influence of disease modifying therapy on radiographic outcome in inflammatory polyarthritis at 5 years: results from a large observational inception study. Arthritis Rheum 2003;48:46-53.