Two knees or one person: data analysis strategies for paired joints or organs

Regression models are being increasingly used in rheumatology because of greater awareness of their application, and the ready availability of computerised statistical packages. Many researchers, therefore, now feel confident to undertake quite sophisticated multivariate analyses. One caution that we feel should be more widely debated and is pertinent to rheumatological datasets in particular, may limit the extent to which these analyses can be routinely undertaken without formal statistical assistance.[1]

The problem stems from the fact that in many aspects of medical research, data are collected on subjects and in most cases analysed using the individual as the basic 'unit' in the analysis. This is appropriate where the data are collected on single organ systems as the unit of the analysis is person specific. In the situation of having data on multiple joints or organs, this approach may not be wholly appropriate. For example, if one were examining risk factors for the knee, some variables, such as injury, will be joint specific rather than subject specific.

One example of this is the study by Doherty *et al*.[2] In this study, factors that might predict progression of knee osteoarthritis were studied and included some which might be considered specific to the knee studied, and some which apply globally to the patient (table 1). As demonstrated in the paper, nearly all the knee specific features described showed a high degree of correlation between the right and left sides. The data could be analysed at the level of either patients or joints (sides). In the case of the former, data regarding knee specific factors have to be sacrificed and, as discussed below, choosing which knee to analyse is problematic. If the data are analysed at the level of the knee, the lack of independence of the data between the knees calls into question assumptions underlying statistical assessments.

It would be incorrect, in this example, to analyse such data without taking the association between knees of the same patient into account. This means the standard methods of logistic and conditional logistic regression cannot be used.[3] Several solutions to this problem have been put forward, many of which were developed for the analysis of ophthalmic data where the same problems arise in the correlation between left and right eyes.[4] A brief overview of each method, together with a discussion of its implications for rheumatology research is given below.

Elementary approaches

Before moving on to more statistically complex methods it is worth considering the simpler and in some ways more naive approaches. The basic problem is the correlation between units from the same person. This issue can, however, be ignored by analysing only one unit per person. The 'worst' or the 'best' side could be chosen or even either side taken at random. In rheumatology an argument can be made for analysing the worst leg/knee as this may have the most functional impact. An argument could equally be made for analysing the best hand as one can compensate to a degree for the poorer one. Analysing the dominant hand would also be another possibility. This approach has the advantage that standard analysis software and techniques can be used. The disadvantage is that not all the side specific data are used. Separate analyses could be carried out on each side to get round this problem but spreading the data over two analyses would mean loss of statistical power—that is, reducing the chances of the study finding significant results—when associations do exist. There is also the added problem of the interpretation (the two analyses clearly are not independent), especially if the results for each side differ.

Table 1 Possible risk factors in a study investigating synovial fluid concentrations of inorganic pyrophosphate and short-term radiographic progression of knee osteoarthritis

Person specific risk factor	*Knee specific risk factor*	*Unclear as to whether person or knee specific*
Sex	Current radiographic severity?	Synovial fluid inorganic pyrophosphate concentration
Age		CPPD crystals

All the above approaches can be considered correct but are restrictive and not efficient; to enter both sides into the analysis without taking the correlation between units from the same person is, however, flawed. There are other more advanced approaches that take this correlation into account,[5] which can be used in different situations and are described below.

Advanced solutions

The problem can be described as one involving correlated covariate data (from the left and right sides of each subject), with a binary outcome variable (for example, diseased/non-diseased). There are basically three broadly defined models, described within the statistical literature, for dealing with this situation. These are referred to as: conditional likelihood, marginal likelihood, and random effects models.[5]

CONDITIONAL LIKELIHOOD MODELS

Perhaps the least useful of these are the conditional likelihood models, which are most commonly used in situations where the same measurement is taken at several time points on an individual (repeated measures data). In this way the inferences are conditional on the previous data. In the situation described herein, with values usually taken simultaneously from two different sources (left and right), this approach is not so appropriate. There is also difficulty in interpreting the results because of the conditionality of the model. This is not to say this situation cannot be modelled through conditional likelihood functions, but in our opinion the approach is difficult to generalise.

RANDOM EFFECTS MODELS

Random effects models estimate the within person variation as well as the between person variation, that is to say allows for the correlation between responses on the same person and produces an estimate of its magnitude. This method can be implemented on commercial software (for example, the logistic-binomial regression for distinguishable data option of EGRET[6] can be used in this situation).

MARGINAL MODELS

In essence, this method treats the correlation between sides of the same patient as a 'nuisance factor', thus it adjusts the

analysis accordingly while never actually estimating the magnitude of correlation. These models also have the advantage that software is becoming increasingly available to implement them, and the results have essentially the same interpretation as conventional logistic regression. The algorithms needed to implement these methods are computer intensive, which is one of the reasons why they are just becoming available. One method developed uses what have been termed, 'generalized estimating equations'[1 7] to produce approximate solutions to the model. Macros have been written for some of the more commonly available large statistical packages including SAS[8] and S+/MIL[9] to implement them. These are available through the internet (addresses given below).

The above overview is intended to be no more than an outline of the methods available to the researcher analysing datasets with correlated binary outcomes. In addition to the greater complexity in applying these models, there are several more detailed issues that need considering when undertaking such an analysis; one of the most important being dealing with missing values. The issue of how the correlation affects statistical power is also pertinent in the design as well as the analysis of a study. Researchers are therefore recommended to consult expert statistical advice on these issues.

In conclusion it is important to appreciate that the problem of non-independence of the units being analysed needs to be considered when analysing data from multiple organs. This problem not only affects multivariate regression analysis, dealt with here, but any other analysis where independence of units is an assumption of the underlying model. This includes the analysis of (unadjusted) 2×2 tables and t tests, where again there is a considerable literature on how the analysis can be adjusted. The strategies suggested above represent solutions that should be considered when such data have been collected and are to be analysed using regression techniques, and careful thought is required to apply and correctly interpret these approaches.

ALEX J SUTTON
KEN R MUIR
ADRIAN C JONES

Department of Public Health and Epidemiology, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH

Correspondence to Dr K Muir.

1 Zhang Y, Glynn RJ, Felson DT. Musculoskeletal disease research: should we analyze the joint or the person? J Rheumatol 1996;23:1130-4.
2 Doherty M, Belcher C, Regan M, Jones A, Ledingham J. Association between synovial fluid levels of inorganic pyrophosphate and short term radiographic outcome of knee osteoarthritis. Ann Rheum Dis 1996; 55:432-6.
3 Breslow NE, Day NE. *Statistical methods in cancer research. Vol 1. The analysis of case control studies*. Lyon, France: International Agency for Research on Cancer, 1987.
4 Rosner B. Multivariate methods in ophthalmology with application to other paired-data situations. Biometrics 1984;40:1025-34.
5 Pendergast JF, Gange SJ, Newton MA, Lindstrom MJ, Palta M, Fisher MR. A Survey of methods for analyzing clustered binary response data. International Statistical Review 1996;64:89-118.
6 Epidemiological Graphics Estimation and Testing (EGRET). Seattle: Statistics and Epidemiology Research Corporation and Cytel Software Corporation, 1985-91.
7 Zeger SL, Liang KY, Albert P. Models for longitudinal data, a generalized estimating equation approach. Biometrics 1988;44:1049-60.
8 SAS Institute Inc. SAS language and procedures: usage. Version 6.1st ed. Cary, NC: SAS Institute, 1989.
9 S+ Manual, Oxford: Statsci Europe, 1995.

Internet addresses for SAS and S+ General Estimating Equation macros
S+ :http://lib.stat.cmu.edu/
SAS :http://statlab.uni@heidelberg.de/statlib/.statlib.html

Recommended further reading
Ashby M, Neuhaus JM, Hauck WW, *et al.* An annotated bibliography of methods for analysing correlated categorical data. Statistics in Medicine 1992;11:67-99.
Rosener B. Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. Biometrics 1982;38:105-14.
Stokes M, Davis CS, Kock GG. Modelling repeated measurements data. In: *Categorical data analysis using the SAS system*. Cary, NC, USA: SAS Institute Inc, 1995: chapter 13.
Williamson J, Kim K. A global odds ratio regression model for bivariate ordered categorical data from ophthalmologic studies. Statistics in Medicine 1996; 15:1507-18.
Zhang Y, Glynn RJ, Felson DT. Musculoskeletal disease research: should we analyze the joint or the person? J Rheumatol 1996;23:1130-4.