STATISTICS IN MEDICINE, VOL. 17, 59–68 (1998)

CHOOSING AMONG GENERALIZED LINEAR MODELS APPLIED TO MEDICAL DATA

J. K. LINDSEY AND B. JONES*

Department of Medical Statistics, School of Computing Sciences, De Montfort University, The Gateway, Leicester LE1 9BH, U.K.

SUMMARY

When testing for a treatment effect or a difference among groups, the distributional assumptions made about the response variable can have a critical impact on the conclusions drawn. For example, controversy has arisen over transformations of the response (Keene). An alternative approach is to use some member of the family of generalized linear models. However, this raises the issue of selecting the appropriate member, a problem of testing non-nested hypotheses. Standard model selection criteria, such as the Akaike information criterion (AIC), can be used to resolve problems. These procedures for comparing generalized linear models are applied to checking for difference in T_4 cell counts between two disease groups. We conclude that appropriate model selection criteria should be specified in the protocol for any study, including clinical trials, in order that optimal inferences can be drawn about treatment differences. © 1998 John Wiley & Sons, Ltd.

Statist. Med., **17**, 59–68 (1998)

1. INTRODUCTION

Many clinical trials and other medical studies involve responses that might be considered to have a normal distribution. However, this is not invariably the case and models based on this distribution are often indiscriminately applied to data which might be better handled otherwise. This is especially true of count data. One possibility is to transform the data to normality, for example by using the approach of Box and Cox.[1] The problems with incorporating this procedure in the protocol for a medical study have recently been discussed by Keene.[2]

An alternative approach which may yield models that are more biologically reasonable in many situations is to use a generalized linear model.[3] Whereas normally distributed data are interpreted as arising as the sum of a large number of unknown factors, Poisson data may be thought, for example, to be produced as counts of random events of constant intensity and gamma data, in certain contexts, as durations that are sums of exponentially distributed times.

In a generalized linear model, the mean is transformed, by the link function, instead of transforming the response itself. The two methods of transformation can lead to quite different

* Correspondence to: B. Jones, Department of Medical Statistics, School of Computing Sciences, De Montfort University, The Gateway, Leicester LE1 9BH, U.K.

CCC 0277–6715/98/010059–10$17.50

Received November 1995
Revised April 1997

results; for example, the mean of log-transformed responses is not the same as the logarithm of the mean response. In general, the former cannot easily be transformed to a mean response. Thus, transforming the mean often allows the results to be more easily interpreted, especially in that mean parameters remain on the same scale as the measured responses.

Other advantages of this family, as compared to normal-based models, include fitting skewed distributions, allowing the variance to be non-constant (varying as a function of the mean), and providing a choice of scales, through the link transformation of the mean, one of which may yield an additive (no interaction) linear model. Thus, various types of data, continuous, binary, or count, can be modelled by members of this richer class of generalized linear models. This does not, however, imply that one should restrict choice to this family. Sometimes, analysis will indicate that a model outside the generalized linear family is required, the most common case being when overdispersion is present in count data, as illustrated below.

Although generalized linear models have been available for almost 25 years, they are not as widely used in medical statistics as might be expected, with two notable exceptions: classical normal models and logistic models. Now that such models are generally available in most statistical packages and not just in the more specialized ones like GLIM,[4] it is important that their advantages be more widely known.

One handicap in choosing a generalized linear model has been their comparison to determine which might best fit a particular data set. For a frequentist, this is extremely difficult because the models are not nested, so that likelihood ratio tests are not applicable. Certain special procedures have been proposed, as by Cox,[5] Atkinson,[6] and Royston and Thompson,[7] but they are rather *ad hoc*, not really fitting in with the mainstream of frequentist theory, this perhaps explaining why they have not become widely used.

However, with the wider use of Bayesian and direct likelihood methods in medical statistics, where the nesting restriction does not apply, this problem does not arise. For example, one simple solution is to use the Akaike information criterion[8] (AIC) or the Bayesian information criterion[9] (BIC) for model selection. As we shall see below, for the relatively small samples often found in clinical trials, they will give similar results; here we shall be primarily concerned with the AIC. As is usual in direct likelihood and Bayesian procedures, we are interested in inferences from a given small sample of data, not with the asymptotic properties of the methods. Frequentist small sample techniques look at the whole sample space, not just the given data.

Recent texts where use of the AIC has been recommended in medical statistics include Clayton and Hills,[10] Collet,[11] and Lindsey.[12] For strong arguments in its favour in one particular biostatistics context, see Burnham *et al.*[13]

When counts are small, a Poisson distribution may often be suitable. Such is not the case for larger counts, where the choice of distribution may be in question. One common example is the use of blood cell counts in medical diagnosis. We shall consider such an example below as a potential application of model selection within the generalized linear model family.

2. MODEL SELECTION

In its general form, a canonical generalized linear model, for one observation, y_i, can be written

$$f(y_i;\theta_i,\delta) = \exp\{[y_i\theta_i - c(\theta_i)]/\delta + d(\delta, y_i)\} \tag{1}$$

where θ_i is the canonical location parameter and δ a dispersion parameter, $c(\cdot)$ and $d(\cdot)$ being known functions. The associated canonical linear regression is

$$\theta_i = \sum_j \beta_j x_{ij} \tag{2}$$

where the x_{ij} are explanatory variables with fixed known values. In fact, the function, $c(\theta_i)$ completely characterizes the model within the family, so that model selection might be thought to involve the choice of an appropriate function here.

For example, the Poisson model can be written

$$f(y_i;\theta_i) = \exp\{y_i\theta_i - \exp(\theta_i) - \log(y_i!)\} \tag{3}$$

where $\delta = 1$ and $\theta_i = \log(\mu_i)$ with μ_i the Poisson mean, so that the function, $c(\theta_i) = \exp(\theta_i)$, characterizes this model.

Usually, the likelihood function is only defined as proportional to the probability (density), so that all parts not involving the parameters can be omitted. Thus, from equation (3), the Poisson likelihood, still for a single observation, would be written

$$L(\theta_i; y_i) = \exp\{y_i\theta_i - \exp(\theta_i)\}.$$

However, to compare models for different members of the family, the likelihood must use the complete probability (density) from equation (1), including all constants not involving the parameters.[14,15] Recall also that a likelihood gives the probability of the data for the given model, thus yielding a criterion for direct comparison among models, those making the data more probable being said to be more likely.

From this likelihood, the AIC for comparing models can easily be derived. It is just based on a penalized likelihood which takes into account the number of parameters estimated in each model. Thus, if we consider the minus two log-likelihood based on equation (1), for the classical AIC, we simply add two times the number (p) of parameters estimated in the model:

$$-2\log[f(\theta,\delta;y)] + 2p = -2\sum_i \{[y_i\theta_i - c(\theta_i)]/\delta - d(\delta, y_i)\} + 2p$$

now for all n observations. The count of parameters, p, will generally include the β_j's in equation (2) plus one for δ, if it is estimated. Because larger (penalized) likelihoods are preferable and we are working on the negative log scale, smaller values of this AIC point to better models. Evidently, a difference of two between two models indicates that the model with smaller of the two AICs could have been one parameter more complex, with the same fit (likelihood) and would still have been competitive, but, if the AIC is at all smaller for one model than another, it is indicating a preference for the former.

For the Poisson models, we obtain, as our AIC

$$-2\log[L(\theta;y)] = -2\left\{\sum_i [y_i\theta_i - \exp(\theta_i) - \log(y_i!)] - p\right\}$$

while, for the normal distribution, it will be

$$-2\log[L(\theta,\delta;y)] = -2\left\{\sum_i [(y_i\theta_i - \theta_i^2)/(2\delta) - y_i^2/(2\delta) - \log(2\pi\delta)/2] - p\right\}$$

where $\theta_i = \mu_i$ is the normal mean and $\delta = \sigma^2$ is the normal variance. Thus, here the function characterizing the model is $c(\theta_i) = \theta_i^2$. Similar equations for the gamma and inverse Gaussian models can also be developed, and will be used below.

When the models to be compared are nested, so that the simpler model is the null hypothesis, the classical AIC, as just described, is, in some ways, analogous to a significance test that is always used with some fixed level. In the same way that this level can be changed, to 1 per cent, 5 per cent, or 10 per cent, more or less stringent criteria can be obtained in the AIC by changing the penalty, that is, the factor that multiplies the number of parameters. Thus, Bhansali and Downham[16] and Atkinson[17] suggest trying more severe criteria, such as a factor of three times the number of parameters. Just as a test level corresponds to a confidence interval, a given AIC penalty corresponds to the precision required of the parameter for it to be eliminated; the greater the penalty, the wider the precision interval. Obviously, this penalty must be fixed before looking at the data, preferably in the protocol.

On the other hand, the BIC, with the factor of two replaced by $\log(n)$, automatically becomes more severe as sample size increases, but, as we can see, provides similar results to the AIC for small n. Bai *et al.*[18] compare various approaches for the special case of log-linear models, which is closely related to Poisson regression, but without comparing distributions as we do here.

Model selection criteria, such as the AIC, were originally designed specifically for prediction purposes in time series (although they are much more widely used now). This is one of the reasons that they do not provide the same results as classical frequentist tests. Thus, for differences of up to seven parameters between two models, the AIC (with a factor of two) will yield more complex models than a test fixed at 5 per cent but it will indicate simpler models than the test for larger differences in the number of parameters. For example, the usual AIC with a penalty of two corresponds to a 15·7 per cent significance level for 1 d.f., but to a 5·1 per cent level for 7 d.f. For a penalty of three, these are respectively 8·3 per cent and 0·4 per cent so that the penalty against more complex models increases rapidly.

One major advantage of proper model selection criteria, such as the AIC and the BIC, as compared to testing (which assumes that hypotheses are specified before inspecting the data) is that they avoid a contradiction of testing, what Lindsey (references 19 and 20, pp. 104–110, 212, 302) calls compatibility. Thus, for example, two orthogonal parameters each tested as non-significantly different from zero at the 5 per cent level (say both with chi-squared values of about 3·5) will simultaneously be shown to be different from zero at the same significance level (chi-squared of 7 with 2 degrees of freedom). Such a compatibility contradiction cannot occur with criteria such as the AIC or BIC.

Once the appropriate penalty has been selected, the AIC will indicate which models are more appropriate, given the data. However, just as one does not use a significance test blindly, always rejecting when the 5 per cent level is attained, so also one will not always select the model with the smallest AIC when several have similar values. Often, there will be good scientific reasons for preferring one model to another and this will be taken into account in the choice. The AIC only provides an objective evaluation of the models given the current data.

3. MODELLING BLOOD CELL COUNTS

Altman (reference 21, p. 199) provides counts of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated

Table I. T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 patients in remission from disseminated malignancies, ordered within groups (reference 21, p. 199)

Hodgkin's disease	Non-Hodgkin's disease
171	116
257	151
288	192
295	208
396	315
397	375
431	375
435	377
554	410
568	426
795	440
902	503
958	675
1004	688
1104	700
1212	736
1283	752
1378	771
1621	979
2415	1252

malignancies, as shown in Table I where they have been reordered within groups to make more evident the form of the distributions. A simple naive approach to model the difference in cell counts between the two diseases would be to look at the difference in estimated means and to make inferences using the estimated standard deviation. Such a procedure implicitly assumes a normal distribution. If we calculate a Student t-test for no difference in means, we obtain a value of 2.11.

Because these are counts, a more sophisticated method would be to use a simple Poisson regression (or log-linear model)

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$$

where μ_{ij} is the mean for the ith individual in the jth group and x_{ij} is an indicator variable, zero when $j = 1$ and one when $j = 2$. Because this model uses logarithms, we are looking at the ratio of the means instead of their difference. Here, the asymptotic Student t-value for $\beta_1 = 0$ is 36·40, quite different from the previous one. Such a result is not surprising, being a result of overdispersion, that is, heterogeneity among individuals within each disease group.

Still a third approach would be to take logarithms before calculating the means and standard deviation, thus, in fact, fitting a log-normal model. In the Poisson model, we looked at the difference in log mean, whereas here we have the difference in mean logs. This procedure gives a Student t-value of 1·88, yielding a still different conclusion.

Let us now use our direct likelihood approach and consider the AICs of these models, as well as some other members of the generalized linear model family. None of the continuous distributions

Table II. Comparison of various distributional assumptions for the T_4 cell count data of Table I, comparing models with no difference in the distribution of cell counts between the two diseases to those having a difference

Model	AIC		Difference in $-2\log(L)$	Estimate/SE
	No difference	Difference		
Normal	608·8	606·4	4·4	2·11
Log-normal	590·1	588·6	3·5	1·88
Gamma	591·3	588·0	5·3	2·14
Inverse Gaussian	590·0	588·2	3·8	1·82
Poisson	11652·0	10294·0	1360·0	36·40
Negative binomial	589·2	586·0	5·2	2·36

in this family is really reasonable *a priori*, but all may be more justifiable than the normal distribution, given the skweness of the data.

These models are presented in Table II. We see, as might be expected with such large counts, that the Poisson model fits very poorly. The other specifically count model, that allows for overdispersion, the negative binomial (not a generalized linear model, but closely associated with them), fits best, while the gamma is second.

The negative binomial distribution can be derived from the Poisson when the mean parameter of the latter is not identical for all members of the population. If this parameter is given a gamma distribution and then integrated out, a negative binomial distribution results; see, for example, Lindsey (reference 12, p. 235). This distribution thus provides one way of modelling heterogeneity in a population.

The choice of the negative binomial, and the clear rejection of the Poisson, indicates that the rate of T_4 cells is not randomly distributed among individuals within a disease group, which would correspond to the Poisson model. Thus, either the rate varies among individuals, over time for each given individual, or among blood measurements on each individual. All are reasonable biological explanations, and the truth is probably closer to a combination of all three factors. To distinguish among them, we would require repeated measurements on the individuals.

Because all of the models in a given column (either difference or no difference) in Table II, except the Poisson, have the same number of parameters, we are in fact comparing log-likelihoods. Thus, for example, with difference between diseases, the negative bionomial makes the data 2·72 ($= \exp[(588{\cdot}0 - 586{\cdot}0)/2]$) times more probable than the gamma distribution while the gamma makes the data 1·35 ($= \exp([588{\cdot}6 - 588{\cdot}0/2]$) times more probable than the log-normal distribution, all models taken at their respective maximum likelihood estimates. In a similar way, comparing across columns, with the negative binomial distribution, a model with disease difference makes the data 13·46 ($= \exp[(589{\cdot}2 - 586{\cdot}0 + 2)/2]$) times more probable than one without difference, but the former model has an advantage of one parameter which is penalized by the AIC. These are *exact* small sample interpretations that require no large sample asymptotic justification.

Traditionally, a frequentist has little means of judging which of these model distributions best fits the data. For example, study of residual plots helps little here because none of the models,

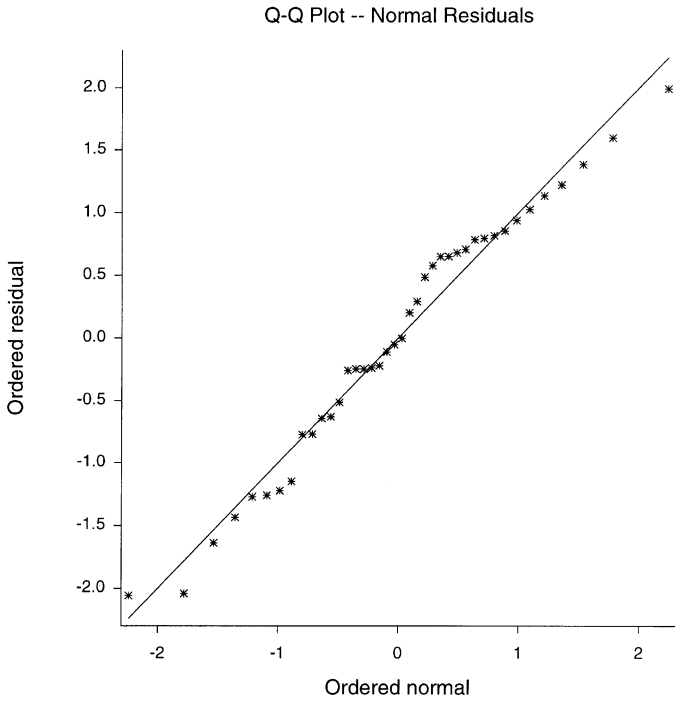

Figure 1. Q–Q plot of deviance residuals for the model using the log-normal distribution with a difference in mean log cell count between the two diseases

except the Poisson, show obvious discrepancies. All of the Q–Q plots are very similar, as illustrated by that for the log-normal distribution in Figure 1.

Let us now look at the difference between the two diseases. From the AIC, a difference is indicated for all distributions. This is not the case if we apply a significance test at the 5 per cent level or a 95 per cent confidence interval, either based on the log-likelihood ratio (the difference in minus two log-likelihood in the second last column of Table II) or the standard error (for example, the Wald test using the ratio of the estimate to the standard error in the last column of the table). On the other hand, if we had chosen the BIC, the factor would be 3·67 ($= \log(40)$) instead of 2, thus here giving results almost identical with a test fixed at 5 per cent.

For non-normal data, tests and confidence intervals based on the likelihood function are more trustworthy than those based on the standard error. Because of the skewness of the distributions, the intervals will not be symmetric. Notice that, in Table II, the log-likelihood and (squared) estimate divided by standard error values for the gamma and inverse Gaussian distributions are considerably different, the likelihood-based tests being more significant. In the table, all values of the $-2\log(L)$ larger than 3·84 indicate that a zero difference between groups will not be included in an asymptotic 95 per cent likelihood-based confidence interval.

The estimated difference in log mean for our best fitting model, the negative binomial, is $-0{\cdot}455$ with standard error, 0·193, indicating lower counts for non-Hodgkin's disease patients. The ratio of mean counts is then estimated to be $\exp(-0{\cdot}455) = 0{\cdot}634$, with an approximate 85

per cent likelihood-based confidence interval of (0·48, 0·84), corresponding to the penalty of two in the AIC. This compares, for example, to an (85 per cent) interval of (0·50, 0·90) obtained from the mean of the ratio of counts assuming a log-normal distribution.

Altman (reference 21, p. 199) also provides the counts of T_8 cells/mm^3 in blood samples from the same 40 patients. Using the same procedures as above, we obtain similar results, with however the log-normal about tied with the negative binomial (difference in AIC of 0·2). The gamma and inverse Gaussian distributions have AICs about two larger. Because of its biological interpretation, discussed above, and the consistency for the two types of counts, we would choose the negative binomial. With these counts, the difference between the two types of diseases is much larger.

For two much more complex examples of this model selection procedure involving cross-over trials, the reader is referred to Lindsey and Jones.[22]

4. DISCUSSION

We have here presented a small sample likelihood interpretation of model selection criteria such as the AIC, and extensions to it. In medical statistics, exact small sample methods are extremely important. As anyone who is accustomed to calculating sample sizes knows, in cases where the calculation is not made and an inordinately large sample used, small clinically irrelevant differences will be detected, while patients are needlessly placed at risk and money wasted. Thus, asymptotic large sample criteria, as usually are used to justify the AIC, are irrelevant in such a context; for the further discussion, see Burnham *et al.*[13] and Lindsey.[20]

In the study of treatment effects or group differences, choice of relevant covariates is usually considered to be an essential part of modelling strategy, at least in exploratory situations. This is one of the simpler and best known model selection problems, generally handled by significance testing to remove non-significant explanatory variables. However, as we have seen here, once we abandon the nesting constraints of frequentist testing, more general model selection, such as among members of the generalized linear model family, is conceptually the same. In such situations, selection of an appropriate model, as described above, should enhance the chances of drawing the correct conclusions about the treatment effect or group difference of interest.

As can be seen from the case of the negative binomial distribution, our methods are in no way restricted to the family of generalized linear models. We have concentrated on them because of their importance and because of the ready availability of software to fit them.

One should here remember that a classical test for treatment difference performed after model selection, whether of covariates or of the distributional assumptions, will not provide the correct significance level because the uncertainty in the model selection process has not been taken into account.[23] At least in simple cases, it is known that model selection to eliminate unnecessary covariates narrows the confidence intervals of the remaining ones.[24]

In the design of clinical trials, the list of pertinent covariates which need to be allowed for in testing for a treatment difference should be specified in the protocol. Preliminary checking whether they are necessary is a model selection problem, preferably performed at the early phases of the study of a medication. In the same way, a protocol should specify that a given class of distributions, such as those in the generalized linear model family, will be considered and the appropriate one selected before reaching a confirmatory trial where the final test concerning treatment difference is to be made.

Generally, such an exploratory process will be performed in phase I and II trials, so that the appropriate model, including distributional assumptions and relevant covariates, should be reasonably well known by phase III. By then, exploratory work should be fairly complete and an appropriate, completely specified, model available, both distribution and covariates; only the final, crucial, confirmatory test of treatment difference should be left. This is the theory, that is, unfortunately, often far from reality. Early trials are usually too small to provide reliable answers, while subsequent trials may even be carried out by a different research group. Nevertheless, only if this procedure is followed will the phase III test of treatment difference have the actual calculated significance level. Of course, secondary analyses of phase III results may also subsequently use exploratory model selection techniques as described here.

In more general contexts than pre-registration development of medicines by the pharmaceutical industry, analysts of confirmatory studies must exercise extreme caution when making conclusions about effects based on exploratory modelling. Model selection criteria such as the AIC are more appropriate than significance tests.

GLIM4 macros for calculating the AIC for generalized linear models are available in Lindsey.[25] Similar procedures can easily be developed for any software package with a generalized linear model option, basing them on the equations given above in Section 2.

ACKNOWLEDGEMENTS

The authors thank the two referees for very helpful comments and suggestions that greatly improved the paper.

REFERENCES

1. Box, G. E. P. and Cox, D. R. 'An analysis of transformations', *Journal of the Royal Statistical Society, Series B*, **26**, 211–252 (1964).
2. Keene, O. N. 'The log transformation is special', *Statistics in Medicine*, **14**, 811–819 (1995).
3. Nelder, J. A. and Wedderburn, R. W. M. 'Generalized linear models', *Journal of the Royal Statistical Society Series A*, **135**, 370–384 (1972).
4. Francis, B., Green, M. and Payne, C. *The GLIM System. Release* 4 *Manual*, Oxford University Press, Oxford, 1993.
5. Cox, D. R. 'Tests of separate families of hypotheses', *Proceedings of the* 4*th Berkeley Symposium*, **1**, 105–123 (1961).
6. Atkinson, A. C. 'A method of discriminating among models', *Journal of the Royal Statistical Society, Series B*, **32**, 323–353 (1970).
7. Royston, P. and Thompson, S. G. 'Comparing non-nested regression models', *Biometrics*, **51**, 114–127 (1995).
8. Akaike, H. 'Information theory and an extension of the maximum likelihood principle', *in* Petrov, B. N. and Csáki, F. (eds), *Second International Symposium on Inference Theory*, Akadémiai Kiadó, Budapest, 1973, pp. 267–281.
9. Schwarz, G. 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464 (1978).
10. Clayton, D. and Hills, M. *Statistical Models in Epidemiology*, Oxford University Press, Oxford, 1993.
11. Collett, D. *Modelling Survival Data in Medical Research*, Chapman and Hall, London, 1994.
12. Lindsey, J. K. *Modelling Frequency and Count Data*, Oxford University Press, Oxford, 1995.
13. Burnham, K. P., White, G. C. and Anderson, D. R. 'Model selection strategy in the analysis of capture–recapture data', *Biometrics*, **51**, 888–898 (1995).
14. Lindsey, J. K. 'Comparison of probability distributions', *Journal of the Royal Statistical Society, Series B*, **36**, 38–47 (1974).
15. Lindsey, J. K. 'Construction and comparison of statistical models', *Journal of the Royal Statistical Society Series B*, **36**, 418–425 (1974).

16. Bhansali, R. J. and Downham, D. Y. 'Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion', *Biometrika*, **64**, 547–551 (1977).
17. Atkinson, A. C. 'A note on the generalized information criterion for choice of a model', *Biometrika*, **67**, 413–418 (1980).
18. Bai, Z. D., Krishnaiah, P. R., Sambamoorthi, N. and Zhao, L. C. 'Model selection for log-linear models', *Sankhyā*, **B54**, 200–219, (1992).
19. Lindsey, J. K. 'The uses and limits of linear models', *Statistics and Computing*, **5**, 87–89 (1995).
20. Lindsey, J. K. *Parametric Statistical Inference*, Oxford University Press, Oxford, 1996.
21. Altman, D. G. *Practical Statistics for Medical Research*, Chapman and Hall, London, 1991.
22. Lindsey, J. K. and Jones, B. 'Treatment-patient interactions for diagnostics of cross-over trials', *Statistics in Medicine*, **16**, 1955–1964 (1997).
23. Chatfield, C. 'Model uncertainty, data mining, and statistical inference', *Journal of the Royal Statistical Society*, *Series A*, **158**, 419–466 (1995).
24. Altham, P. M. E. 'Improving the precision of estimation by fitting a model', *Journal of the Royal Statistical Society*, *Series B*, **46**, 118–119 (1984).
25. Lindsey, J. K. 'The AIC for comparing models in GLIM', *GLIM Newsletter*, **25**, 6–8, (1995).

