Likelihood-Based Diagnostics for Influential Individuals in Non-Linear Mixed Effects Model Selection

Sima Sadray,[1,2] E. Niclas Jonsson,[1] and Mats O. Karlsson[1,3]

Received March 10, 1999; accepted April 15, 1999

Purpose. Data from single individuals, or a small group of subjects may influence non-linear mixed effects model selection. Diagnostics routinely applied in model building may identify such individuals, but these methods are not specifically designed for that purpose and are, therefore, not optimal. We describe two likelihood-based diagnostics for identifying individuals that can influence the choice between two competing models.

Methods. One method is based on a jackknife of the raw data on the individual level and refitting the model to each new data set. The second method is a calculation which utilises the contribution each individual make to the objective function values under each of the two models. The two methods were applied to model selection during analysis of a real data set.

Results. The agreement between the methods was high. Individuals for whom there was a discrepancy between the methods tended to be those for which neither of the contending models described the data appropriately. Both methods identified individuals that influenced the model selection.

Conclusions. Two objective, specific and quantitative methods for identifying influential individuals in nonlinear mixed effects model selection have been presented. One of the methods doesn't require additional model fitting and is therefore particularly attractive.

KEY WORDS: model selection; mixed effects modeling; jackknife; case deletion; NONMEM.

INTRODUCTION

Models that describe the pharmacokinetics and pharmacodynamics in the target patient population are often characterized using sparse data and/or non-linear-mixed effects modeling. Different model building strategies have been suggested to appropriately describe the characteristics (1–4), but all involve testing a multitude of models for their fit to the data. The models are often partly tested in a sequential order, where models are compared and the most suitable model, as judged by some criteria, is retained for the next step. Each step may involve the selection of some part of the structural, statistical or covariate model.

It is a recognized risk that the selection of one model over another may heavily rely on a single or a few individuals' data. Individuals may influence model selection for different reasons. First, they may be representative of the population, but contain information that is lacking or sparse in other individuals. There may, for example, be more data from some individuals than others, or their data may be located at times or doses where there is little other data. Also, there may be individuals who are at the extremes of covariate distributions and therefore provide much of the information about the strength of the parameter-covariate relationship. Second, there may be individuals who are truly different from the main population and their data reflect this. Last, there may be error in some of the data and this may be manifested as influential individuals. The first type of influential individual can often be discerned from the last two types, which are usually termed outliers. For outliers, it is often difficult to judge whether they represent a true feature of the individual or error in data. It should be noted that outliers may be without influence on the model selection, but such outliers are not a concern in this presentation. The action taken when an influential individual has been identified depends on the nature of its influence, the purpose of the analysis and whether it is judged to reflect true characteristics, be high leverage or outlying. Regardless of what the interpretation is, knowledge that the model, or a part of it, is influenced by a single or a few individuals is usually sufficient to warrant caution in extrapolation of the results.

Influential individuals may be of importance in at least two respects: selection between models, and, on the components (parameter estimates) of the model(s) of interest. Previous work in non-linear mixed effects modeling has focused on the latter, where Mandema *et al.* (5) suggested a case-deletion strategy to assess the presence of influential individuals. In the building of population pharmacokinetic and pharmacodynamic models, it is common to use individual parameter estimates in preliminary covariate model building (1). Stepwise building of generalized additive models (GAM) have been suggested as a strategy to identify candidate covariate relationships (3). Measures for individual influence on components of the final generalized additive model can be obtained analytically and, in addition, bootstrap approaches to identify influential individuals in the covariate model building have also been described (6). Some of these procedures have been incorporated into the population model building software Xpose (7). However, these procedures rely on the quality of the individual parameter estimates and measure the influence in the nonlinear mixed effects modeling only in an indirect way. Also, they are only applicable for covariate model building and not when influential individuals for structural or statistical model components are to be identified. In the present work, we describe procedures for identifying influential individuals during nonlinear mixed effects model building. We compare an approach based on case-deletion to one based on individual contribution to the likelihood value.

METHODS

Case-Deletion

During model building using the NONMEM program, model selection is typically based on the objective function value, which is approximately minus twice the log likelihood, and as this value is central to the present work, we will elaborate

[1] Division of Biopharmaceutics and Pharmacokinetics, Department of Pharmacy, Faculty of Pharmacy, Uppsala University, Box 580, SE-751 23 Uppsala, Sweden.

[2] Division of Biopharmaceutics and Pharmacokinetics, Department of Pharmacy, Faculty of Pharmacy, Tehran University of Medical Sciences.

[3] To whom correspondence should be addressed. (e-mail mats.karlsson@biof.uu.se)

some on its notation. Normally, the objective function value would be defined by two entities: the model and the data used to fit the model. In the following, we will also concern ourselves with the objective function value of a model applied to data other than those used to estimate its parameter values. Thus three components are necessary to define the objective function value: the structure of the model, the data used to define its parameters, and the data to which the model is applied in the evaluation of the objective function value. Comparisons will be made between a basic (B) and a full (F) model and between data sets that contain the data from all (n) subjects, all but one (n-i) subjects or a single (i) subject. To exemplify the notation, the objective function value of the basic model with parameters derived from all data and applied to the data from subject i will be denoted $O_{B,n,i}$.

One way of assessing the influence of a single individual is to monitor how the difference in objective function value between two models changes when it is based on all data compared with all data minus the individual in question. $\delta O_{jackknifc.i}$ given by Eq. 1, will provide such a diagnostic for identifying individuals whose data support the more complex model and individuals whose data will not. The former will have negative values and the latter zero or positive values. Individuals with high positive values of $\delta O_{jackknifc.i}$ will be referred to as 'masking' individuals, as without their presence, the full model would provide a larger improvement in the description of the data. Conversely, individuals with high negative values of $\delta O_{jackknifc.i}$ will betermed 'driving'.

$$\Delta O_{jackknife.i} = (O_{F,n,n} - O_{B,n,n}) - (O_{F,n\text{-}i,n\text{-}i} - O_{B,n\text{-}i,n\text{-}i}) \quad (1)$$

A straight-forward procedure to obtain $\Delta O_{jackknife,i}$ is to fit both the full and reduced model to the full data set and to a data set where the data of the subject in question has been omitted. $\Delta O_{jackknife,i}$ is then obtained according to Eq. 1. As indicated by the notation, this general methodology is termed the jackknife (8) and although it has not to our knowledge been applied in this particular context, it is a straightforward application of the jackknife.

The objective function value of a model for a given data set is the sum of the individual contributions, where the latter can be obtained as part of the output from a NONMEM analysis. A second procedure that makes use of these and calculates individual contributions to the objective function value difference between the full and reduced model, is shown in Eq. 2. The middle part of the equation is included to demonstrate the similarity between the $\Delta O_{jackknifc,i}$ and $\Delta\ O_{diffcrencc.i}$ measures.

$$\begin{aligned}\Delta O_{diffcrence,i} &= (O_{F,n,n} - O_{B,n,n}) - (O_{F,n,n-i} - O_{B,n,n-i}) \\ &= (O_{F,n,i} - O_{B,n,i}) \quad (2)\end{aligned}$$

$\Delta O_{diffcrence,i}$ obtained in this manner, is termed the "difference" method, since its calculation is a straightforward difference under the same model. The jack-knife and difference procedures will be compared with respect to identification of influential individuals. For illustration, consider a situation where, using the full data set, there is no difference in objective function value between the full and the basic model ($O_F = O_B$). If a certain individual i was omitted, however, a difference between the two models would appear, such that $O_F < O_B$. Such an individual would have been masking the relationship in the full data set, and both $\Delta O_{difference,i}$ and $\Delta O_{jackknifc,i}$ would be positive.

Data

Data from a Phase II, multi-center, dose-finding study of oral moxonidine tablets versus placebo in patients with congestive heart failure are used to assess and illustrate the procedures above. The study was a parallel group design where patients received placebo or one of three moxonidine treatments. Active treatment started at 0.1 mg twice daily and was escalated to a predefined dose, 0.1, 0. 2, or 0.3 mg twice daily. The pharmacokinetic data have been used previously to illustrate methodological work regarding nonlinear mixed effects model building (9). Pharmacokinetic sampling was performed after the first dose and after 12 weeks of therapy, of which the last 8 weeks had been on the same dose. The target sampling times were 0.5, 1, 1.5, 2, 4, 6 and 8 hours after the morning dose. In addition, a trough sample was also collected immediately before intake of the steady state dose. In total, 1022 moxonidine concentration measurements were available from 74 patients. Four subjects were studied only after the first dose as they dropped out of the study before the second study occasion. With respect to the dosing history, patients were assumed to adhere to the twice daily administration with twelve hour dosing intervals. The last dose before the monitoring of the concentration-time profile was administered in the presence of the clinic staff.

Covariates considered during the course of the population analysis presented here were: age (median 66 yrs, range 43–78 yrs), creatinine clearance (CRCL; median 65 ml/min, range 30–142 ml/min) which was calculated according to Cockroft and Gault (10), gender (59 males) and concomitant medication with ACE-inhibitors (present in 47 subjects) or digoxin (present in 48 subjects) as two separate covariates. Other characteristics of the patients are described in Karlsson *et al.* (9).

Modeling

The starting model for describing the data was a one-compartment model with first-order absorption and lag-time. It was defined by the parameters clearance (CL), volume of distribution (V), absorption rate constant (ka) and lag-time (Tlag). For all parameters interindividual variability was also estimated. The observed concentrations were log-transformed and an additive residual error model was used (which approximately corresponds to a proportional error model on the untransformed scale). All parameters of this model have been reported earlier (9). Extensive model building diagnostics suggested that the final model was similar to starting model with the addition that a linear relationship between CL and creatinine CL should be included in the model. In the following some model comparisons will be retested to investigate whether the model choice may have been affected by one, or a few, influential individuals. All modeling was performed using the first-order (FO) algorithm implemented in NONMEM, version V (2). All model comparisons that are reported herein are between hierarchical models, and, for these, the likelihood ratio test is applicable (2). For a one parameter difference between a basic and full model, the difference in objective function values for achieving statistical significance of $p < 0.05$, 0.01 and 0.001 are 3.84, 6.63 and 10.83, respectively. $O_{B,n,n}$ (or $O_{F,n,n}$, $O_{B,n-i,n-i}$ or $O_{F,n-i,n-i}$) is standard output from NONMEM, whereas $O_{B,n,i}$ (or $O_{F,n,i}$) is not. The latter can be obtained by evaluating (without estimation) the data for the single individual i under the final parameter estimates. This can be implemented in a number of different

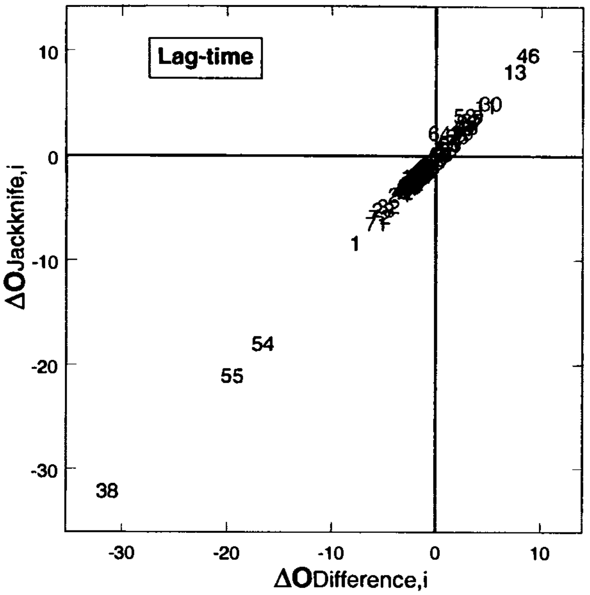

Fig. 1. Identification of influential individuals for selecting a model with or without lag-time. Positive and negative values indicate masking and driving individuals, respectively. Plot symbol is the ID number.

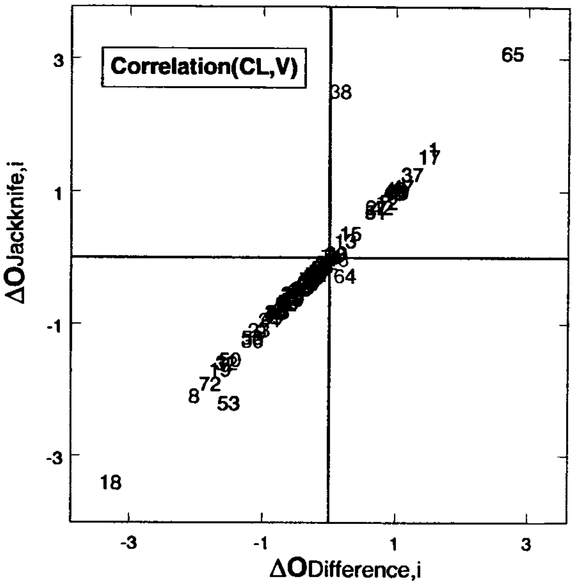

Fig. 2. Identification of influential individuals for selecting a model with or without correlation between CL and V. Also see legend to Fig. 1.

ways into NONMEM. We used a code that extracted these values during a final pass through the data after the population parameters had been determined.

RESULTS

Omitting the lag-time component from the starting model resulted in an increase of the objective function value of 96.4. From Fig. 1 it can be seen that subjects 38, 54 and 55 may be driving this relationship. As expected, $\Delta O_{\text{difference.i}}$ and $\Delta O_{\text{jackknife.i}}$ are similar but not identical. Omitting the three driving subjects and refitting the full and reduced (i.e. with lag-time) models results in decrease in the difference in objective function of 33.0. Thus, the three individuals appear to be responsible for a large portion of the difference between the models with and without lag-time. However, the inclusion of lag-time is not solely dependent upon these individuals, but is also motivated in their absence.

Addition of covariance between CL and V resulted in a decrease in the objective function value of 18.7. Potentially masking (ID 65) and driving (ID 18) subjects can be identified in Fig. 2. Omitting either of these two subjects does not change the model building choice or the parameter estimates in any great way. Subjects 38 and 64, that don't adhere to the general trend of a high correlation between the methods, are further discussed below.

Significant ($p < 0.001$) relationships were found between CL and both CRCL and AGE (Table I; Runs 4 and 5). CL was predicted to be positively and negatively correlated to CRCL and AGE, respectively. As the differences between the reduced and the full models were more than 15, and no single driving individual showed a $\Delta O_{\text{jackknife.i}}$ or $\Delta O_{\text{difference.i}}$ lower than -3.2 (Figs. 3 and 4), it is clear that the result is not dependent on any one single individual. In general, the detected masking and

Table I.

Run	Model	ΔO[a]	Comment
1	Starting	—	
2	−Lag-time	96.4	
3	+Correlation(CL,V)	−18.7	Correlation(CL,V) = 0.59
4	CL ~ CRCL[b]	−18.6	CL (L/h) = 17.7 + 0.14*CLCR
5	CL ~ Age	−16.9	CL (L/h) = 54.5 − 0.42*Age
6	CL ~ ACE-inhibitors	− 7.6	$CL_{ACE}/CL_{no\ ACE} = 1.18$
7	CL ~ Digoxin	− 2.9	$CL_{Digoxin}/CL_{no\ Digoxin} = 1.11$

[a] Objective function value relative to that of Run 1
[b] The sign ~ denotes "is a function of."

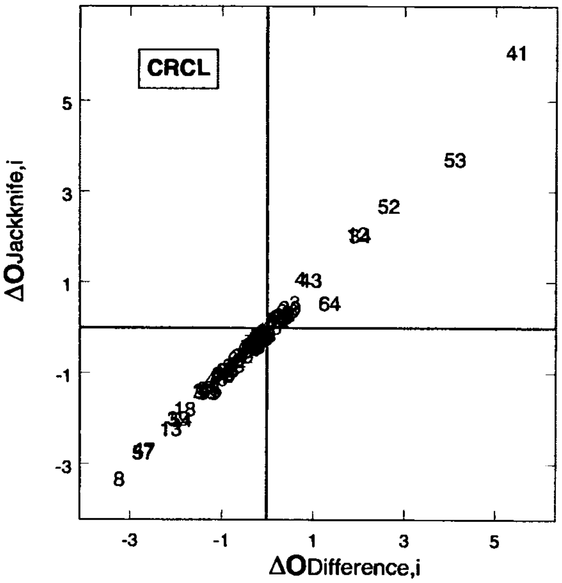

Fig. 3. Identification of influential individuals for selecting a model with or without CLCR influencing CL. Also see legend to Fig. 1.

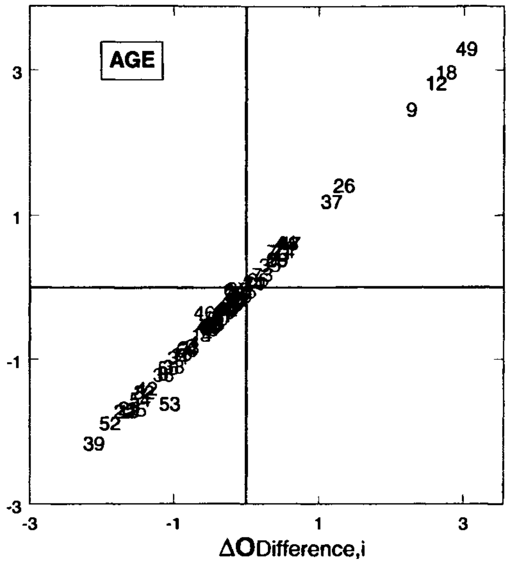

Fig. 4. Identification of influential individuals for selecting a model with or without age influencing CL. Also see legend to Fig. 1.

driving individuals are those that could have been predicted. Thus, for the CL versus AGE relationship, ID's 49, 18 and 12 are young individuals with below average CL, whereas ID 9 is an old subject with above average CL.

Patients on concomitant medication with ACE-inhibitors had a significantly ($p < 0.01$) higher CL than other patients. In a data set where the data from the driving individuals 53, 43 and 48 (Fig. 5) have been omitted, the covariate is no longer significant even at the 0.05 level.

No significant relationship was found between CL and digoxin use (Table I, Run 7). However, individual 13 appears to be a potentially masking subject (Fig. 6) and when this individual was dropped from the data set, a significant ($p < 0.01$) relationships was found.

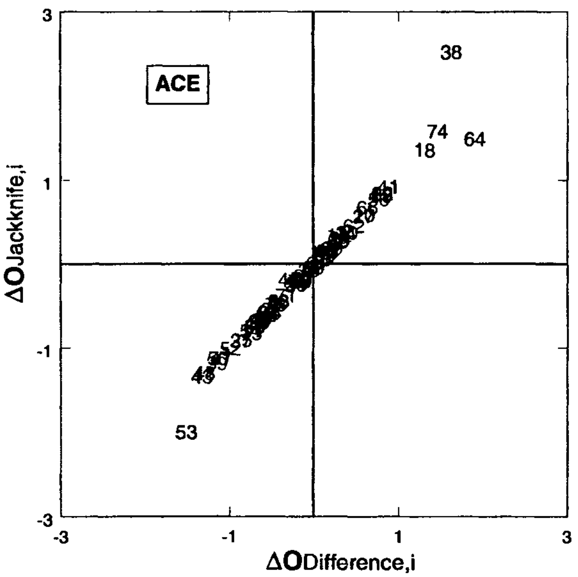

Fig. 5. Identification of influential individuals for selecting a model with or without concomitant medication with ACE-inhibitors influencing CL. Also see legend to Fig. 1.

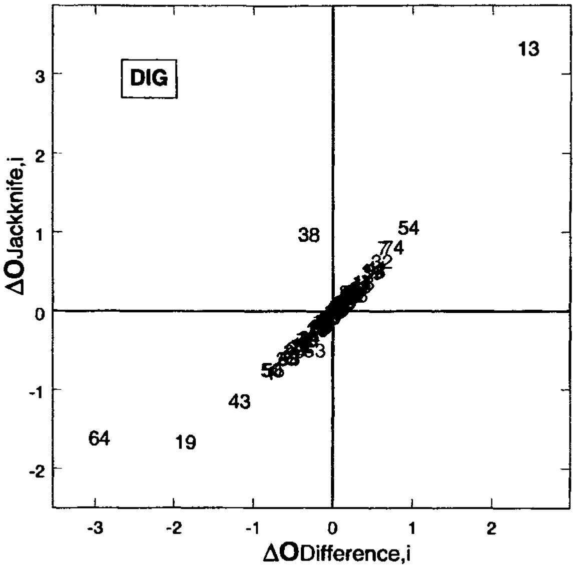

Fig. 6. Identification of influential individuals for selecting a model with or without concomitant medication with digoxin influencing CL. Also see legend to Fig. 1.

It is not surprising that patients with higher CL (ID 8, 13, 25) or lower CL (ID 18, 43, 48), are often influential individuals. Neither is it surprising that individuals with extreme values of continuous covariates, such as the oldest (ID 39, 44, 52) and youngest (ID 12, 49, 55) or those with the highest (ID 8, 54, 72) or lowest (ID 23, 42, 53) CLCR often are among the more influential for the detection of the covariate relationships. There are other individuals that appear not only to be influential individuals in several of the model selections, but also display differences between $\Delta O_{jackknife,i}$ and $\Delta O_{difference,i}$. Most pronounced is this in subjects 38 and 64, but also in subjects 13 and 53. These individuals have higher than average, but not extreme, values of CL. The common feature of these individuals is that their concentration-time profiles are not well characterised by the one-compartment, first-order absorption model. It is notable that the influence of such subjects on model selection is not easily predictable. ID 64, for example, has a high CL and is concomitantly treated with ACE, but despite this appears as a masking individual for this relationship. ID 13 displays a clear lag-time, but yet is masking with respect to the inclusion of a lagtime. ID 38 has likewise a high CL and is concomitantly treated with DIG, but appears as masking as judged by $\Delta O_{jackknife,i}$. For this individual $\Delta O_{difference,i}$ appears to be more in line with what can be expected. The same is true for the difference between two methods with respect to ID 38 and correlation between CL and V (Run 2). The individual estimate of V for this individual was close to that of the typical individual, 115 versus 121 L, respectively. Therefore it seems surprising that the influence of this individuals data on whether a correlation between CL and V exist should be large as indicated by $\Delta O_{jackknife,i}$.

DISCUSSION

The jackknife and difference methods, as implemented in this work, appear to have properties that make them suitable for identifying individuals with a large influence on model selection. The two methods differ from each other in an important aspect. The difference method is used to investigate the influence of an individual assuming that the individual's data can be appropriately described by at least one of the two contending models. The jackknife procedure does not make this assumption. Therefore, it is not surprising that the two methods can give rise to different results and which happens mainly for subjects for whom neither model appears to be ideal. For example, a subject who, under the model(s) in question, can be classified as driving by the difference method, because the parameter and covariate values are both supporting the mean population trend, may be classified by the jackknife method as masking, since the covariate relationship appears stronger when the overly noisy data from the subject is omitted from the data set.

Despite the occasional differences between the methods, the most striking feature is the similarity in the classification of individuals. This indicates that, to a large extent, the two methods could substitute for each other. The difference method has the advantage that all the values necessary to calculate it can be generated as part of a normal run and therefore it can be made available without any extra computational burden. The jackknife method, on the other hand, requires re-estimation of the model parameters as many times as there are subjects in the data set. In addition to the computational burden this causes, there is the problem of unsuccessful terminations of the estimation, or, of termination in a local minimum. In the example presented, approximately 5% of all data sets had to be rerun one or several times for these reasons.

The present investigation was restricted to only study hierarchical models. It is, however, straightforward to adapt these measures to non-hierarchical models if selection criteria based on the objective function value are used. Also, the application has been directed towards model building using the NONMEM program, but the methods should be equally well suited for use with other programs where model selection criteria based on likelihood differences are applied.

Individual contributions to the objective function value may have other uses than those described here. A related application could be their use in determining which subjects influence parameter estimates in sensitivity analyses. Also, the inspection of the individual contributions under a specific model may give insight into which subjects are potentially influential and different from the study population as a whole (Fig. 7). For a data set like the present, where there is little variability in number of observations per subject, subjects with a poor fit can be identified as those with high values. For example, the subjects with highest individual objective function values under the starting model are ID 13, 38 and 64. These subjects were also identified as influencing the model selection in a number of comparisons as discussed above. If bootstrap methodology (8) is to be used for calculation of posterior power (here how frequent a certain feature would be identified in selection between models), a bootstrap of the individual differences may be a time-saving alternative to performing multiple analyses of bootstrap data sets.

There are several routines available for identifying influential individuals in covariate model building. There are also diagnostics that exist that may identify influential individuals during structural and statistical model building, although the power of these general model building tools to identify influential individuals in nonlinear mixed effects model selection has not been specifically addressed in the literature. Such methods are qualitative and are based on empirical Bayes estimates or rely on plots of data, predictions or residuals. The new methods described herein are additional tools for identifying outliers in model selection. Their advantage lies in that they are objective, specific and quantitative. The addition of these diagnostic tools may be particularly useful for identifying influential individuals in the choice of structural and statistical models, where alternative diagnostics are not as applicable as for covariate models. Neither of the two described methods can be practically implemented in model building unless the relevant statistics can be obtained in an automated fashion. Routines for doing so are available from the authors (niclas.jonsson@biof.uu.se).

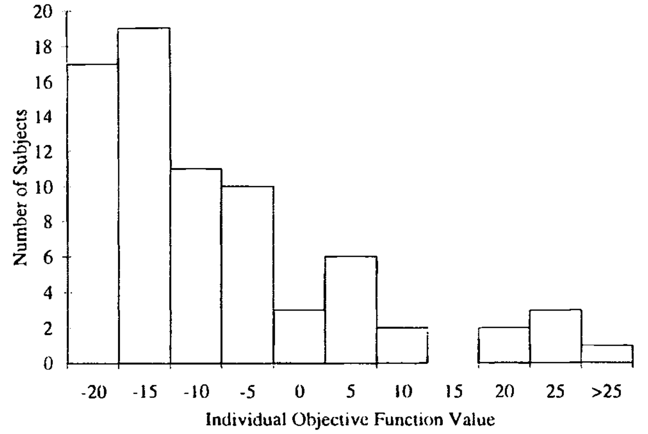

Fig. 7. Histogram of individual contributions to the objective function value for the starting model.

It is clear from the example given that data from a single individual alone can give rise to a difference in objective function value between two models that is highly significant, at least when, as in this case, a relatively rich sampling scheme has been used. At least three negative consequences can be identified by falsely believing a model that is driven by one or a few individuals: (i) the precision in future predictions will be lower than otherwise, (ii) other, real, relationships may be masked, and (iii) collection of information that is unnecessary (spurious covariate effects) or sub-optimal (sampling times) may seem indicated. The opposite, that individuals mask real relationships may be particularly frequent with relationships that originally are based on a relatively few number of subjects, as is often the case with for example drug interactions or when the population study is based on relatively few individuals. The latter is often the case in studies in children, when both number of samples and subjects are generally low.

The suggested diagnostics identify the most influential individuals in model selection. Talthough the scale of the individual contribution is directly related to that used for testing significance between models, the related question of whether the subject is an outlier or only have highly informative data cannot be decided based on this diagnostic alone. Neither can the question of what the appropriate procedure is, once an influential individual has been identified. To quote from a recent overview of diagnostic tools: ". . . model adequacy (a) should be carried out with respect to the substantive questions of interest and (b) cannot be carried out in isolation of the context; in particular the interpretation of diagnostics requires subject-matter knowledge" (11).

The art of population analysis has been increasing in complexity as more tools and diagnostics are developed that aid the model building process. However, this can only be to the good as the data analyst is guided to look more and more closely at the results obtained, their reliability and their predictive capacity. Also, with increased automated diagnostic tools this can be achieved without an increased total analysis time. The new methods presented in this paper should help identify individuals that may drive or mask decisions during the model building process, but should not be considered as a substitute for basic background knowledge that should help the analyst determine if the relationships and models found are fundamentally reasonable.

ACKNOWLEDGMENTS

This work was funded by a grant by Pfizer Central Research, Sandwich, UK.

REFERENCES

1. P. O. Maitre, M. Buhrer, D. Thomson, and D. R. Stanski. A three-step approach combining Bayesian regression and NONMEM population analysis: application to midazolam. *J. Pharmacokinet. Biopharm.* **19**:377–384 (1991).
2. S. L. Beal and L. B. Sheiner (Eds) NONMEM Users Guides. NONMEM Project Group, University of California at San Francisco, San Francisco, 1998.
3. J. W. Mandema, D. Verotta, and L. B. Sheiner. Building population pharmacokinetic-pharmacodynamic models. I. Models for covariate effects. *J. Pharmacokin. Biopharm.* **20**:511–528 (1992).
4. E. N. Jonsson and M. O. Karlsson. Automated covariate model building within NONMEM. *Pharm. Res.* **15**:1463–8 (1998).
5. J. W. Mandema, D. Verotta, and L. B. Sheiner. Building population pharmacokinetic-pharmacodynamic models. In Advanced methods of pharmacokinetic and pharmacodynamic systems analysis. Volume 2. Ed. DZ D'Argenio, Plenum Press, New York, 1995.
6. E. I. Ette and T. M. Ludden. Population pharmacokinetic modeling: the importance of informative graphics. *Pharm. Res.* **12**:1845–1855 (1995).
7. E. N. Jonsson and M. O. Karlsson. Xpose-An SPLUS based population pharmacokinetic-pharmacodynamic model building aid for NONMEM. *Comput. Methods Programs Biomed.* **58**:51–64 (1999).
8. B. Efron and R. J. Tibshirani. An introduction to the bootstrap. Chapman & Hall, New York, 1993.
9. M. O. Karlsson, E. N. Jonsson, C. G. Wiltse, and J. R. Wade. Assumption testing in population pharmacokinetic models: illustrated with an analysis of moxonidine data from congestive heart failure patients. *J. Pharmacokin. Biopharm.* **26**:207–46 (1998).
10. D. W. Cockroft and M. H. Gault. Prediction of creatinine clearance from serum creatinine. *Nephron* **16**:31–41 (1976).
11. J. Wakefield (discussant). In J. S. Hodges. Some algebra and geometry for hierarchical models, applied to diagnostics. *J. R. Statist. Soc. B* **60**:497–536 (1998).