Imputing Missing Data in Clincal Pilot Studies

Monnie McGee* & Nora Bergasa†

*Department of Statistical Science, Southern Methodist University
Dallas, TX 75275

†Division of Hepatology, SUNY Downstate Medical Center
Brooklyn, New York

Abstract

Pilot studies are experiments which typically involve fewer than 20 subjects in order to test the feasibility of a new treatment. Aside from the problems in dealing with a small number of subjects, some of the observations may be missing. In a trial that is already small, one does not want to discard any data and therefore decrease further the efficiency of any estimates. Other issues, such as outliers and detection limits, are important to consider, as well. This paper gives a description of how these issues were resolved in a small clinical trial of the drug gabapentin for treatment of severe scratching in liver disease. Particular attention is paid to imputation of missing data, and a simulation study conducted showing that the chosen imputation method has good statistical properties.

Key Words: mixed-effects models, imputation, hot-deck, longitudinal data

1 INTRODUCTION

Real-life data are hardly ever as clean as textbook examples. This fact is particularly true where humans are the subjects of an experiment. Real data are often unbalanced with missing observations, and they contain a variety of other problems that are not typically discussed in statistics courses. Missing values in a trial that is already small are of particular concern, since one does not want to discard any data and therefore decrease further the efficiency of any estimates.

This work was motivated by a small study on the efficacy of a drug called gabapentin in reducing scratching, secondary to itching, which is a complication of liver disease [1]. Sometimes, the scratching can be so severe that patients cannot sleep. Some patients tear their skin, causing wounds that can be secondarily infected. Accordingly, finding a drug that ameliorates the itching and scratching is an important research effort. Liver disease, particularly in its advanced stages, is very difficult to cure. However, it is hoped that drugs such as gabapentin can make patients more comfortable by reducing the effects of its complications.

Although every effort was made to ensure that subjects kept appointments and complied with the protocol (described in more detail in the next section), many of the observations in the trial were missing. Two subjects dropped out of the trial altogether, and others did not complete all of the required measurements. The observations tend to be missing in chunks (three or four observations missing in a row). In addition, many of the observations contain large outliers which makes model fitting difficult. However, the data, with all its faults, represent three years of work for the physicians, and a significant time investment for the research subjects. Thus, it is important to glean all information possible from the data, in a principled and reasonable way.

This paper tells the story of the preparation of the data from a small clinical trial for analysis with a mixed–effects model. The problems were not in the analysis itself, but in preparing the data for analysis. The issues of small numbers of subjects, missing data, outliers, and detection limits which are discussed pertain to a wide variety of medical studies. The study protocol is described in detail in Section 2. In Section 3, issues in the replacement of missing data for this clinical trial are outlined. A simple and reasonable method for imputing missing observations is described in Section 4. Section 5 describes a secondary issue of detection limits. A mixed–effects model is applied to the imputed data in Section 6. Simulations for power and size for various scenarios involving missing observations replaced in the way described in Section 3 are given in Section 7. The final section presents a discussion of results.

2 THE GABAPENTIN TRIAL

The study was approved by the Investigational Review Board of Columbia University, where the study was conducted. All subjects signed an informed consent prior to participation in the trial. The protocol called for sixteen subjects to be randomized to either gabapentin or a placebo. Before the treatments began, baseline data of scratching activity and perception of pruritis were obtained. Scratching activity was measured by a scratching activity monitoring system that consists of a piezo film sensor glued to a cast custom made to fit the middle finder on the dominant hand of the user [2].

The main component of the system is a signal processor, which consists of a frequency

counter incorporating a threshold detector and a bandpass filter to prevent extraneous counts from being registered. The threshold level was adjusted to allow for approximately 90% of the scratching signal to be captured. At the start of the recording sessions, subjects were asked to scratch a defined distance over a defined surface to yield a counter reading of 2000 to 3000 for a period of 30 seconds.

The counts are added and presented on the print out as hourly scratching activity (HSA). The result is a numerical value which purports to measure the amount of scratching for each subject. Large values indicate more scratching. HSA was measured over a 48-hour period in the hospital. Therefore, the raw HSA data consist of 96 hourly measurements (pre and post measurements over 48 hours) for each subject. When the activity does not meet the scratching threshold, the printed numbers are negative and defined as background movement.

After the initial quantification, subjects were given their randomly assigned medication and asked to resume their normal daily routines. After six weeks on the study medication, the subjects returned to the hospital for a second 48-hour evaluation which was conducted in the same way as the first.

3 ISSUES IN PREPARATION OF THE DATA

Statisticians recognize that real data are usually not well-behaved, and that properly preparing the data is essential in obtaining a valid and reliable statistical analysis. A first step in such preparation is exploring the data with various graphical tools and descriptive statistics. It may also be helpful to examine the data file itself (in whatever software format, i.e. Excel) for potential problems. For example, in the gabapentin study, one value was recorded as 4,85 rather than 4.85, something which would not have been discovered without examination of the Excel file containing the data.

For data taken at regular intervals over a fixed length of time, time plots are a good tool examining the profile for each subject. Figure 1 shows the hourly HSA sequences for each subject in the gabapentin and placebo groups at baseline. Plots of the other two groups post-treatment have similar characteristics. Three things are immediately apparent. First, although it makes sense that there should be correlation structure within patients, any such structure is difficult to visualize due to the presence of very large outliers. Second, some subjects have very low values of HSA (some values are identically zero), and third, not all of the subjects completed the study. In comparing the two treatment groups, one can also see that the HSA levels for the placebo group are typically lower than those of the gabapentin group. Such a discrepancy can be expected, even when randomization is performed correctly, due to the small number of subjects.

For almost any parametric statistical analysis chosen for these data, stable model parameter estimates would be very difficult to obtain due to the extreme outliers. Given the nature of these data, there should be statistical evidence of correlation within subjects for the HSA measurements. In fact, plots of autocovariance and partial autocovariance estimates (not shown) for the HSA measurements for each subject indicate that only two of the subjects exhibit any correlation structure in their HSA measurements. Plots for the other subjects are typical of white noise. Since outliers are the likely cause of this problem one obvious solution is to delete them. Before deletion of outliers, percentages of missing observations vary from 100% (missing entire pre or post measurements) to 2% (missing only one value) for the other fifteen subjects in the study. There are four subjects for whom

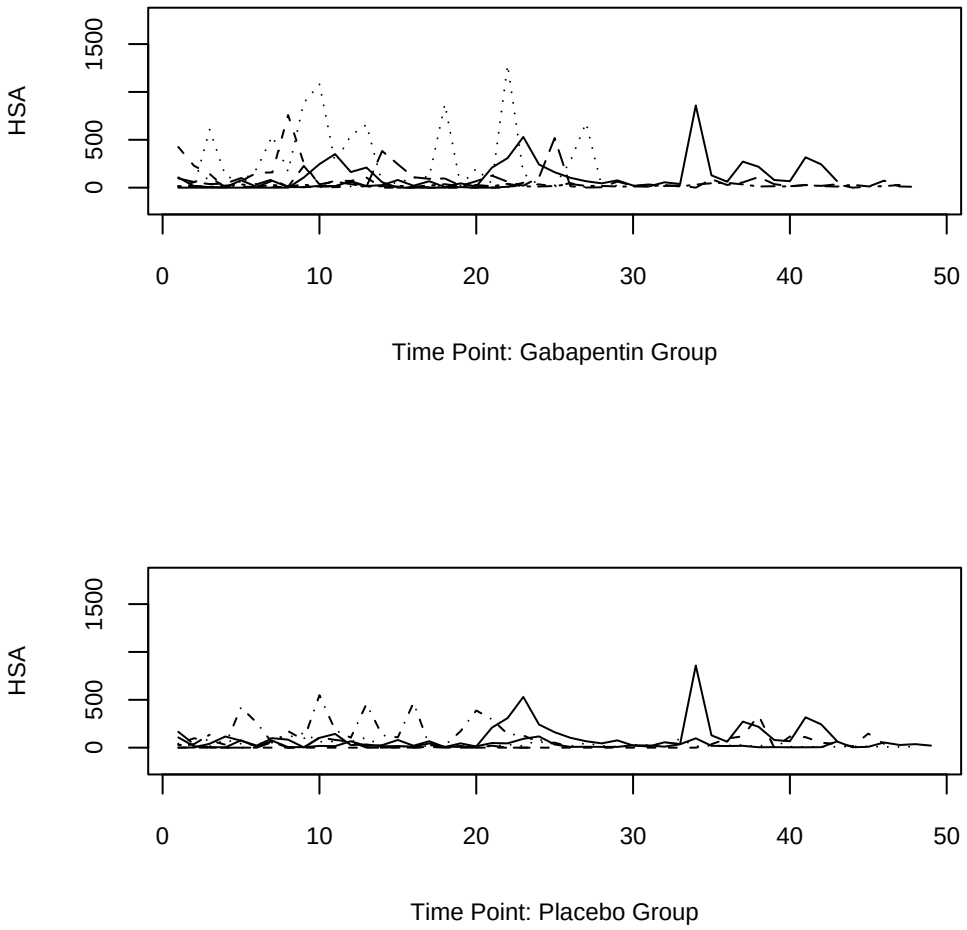

Figure 1: Time plot of pre treatment hourly HSA for the gabapentin (top) and the placebo (bottom) groups.

entire pre- or post-treatment quantifications are missing. For such a small study, deletion of four subjects, in addition to the outliers, would seriously compromise the power of any analysis. A more promising alternative would be to replace all missing observations (as well as outliers) in a principled way.

Although many articles on dealing with missing data pretend that data are missing completely at random (MCAR), most of the time, the true mechanism is really MNAR [3]. The missing observations from the gabapentin trial are generated by a mixture of mechanisms. For some values, we know that the mechanism generating missing values is independent of the observations themselves. For two of the subjects who are missing entire pre or post quantifications, the reason for the failure to collect data was that the machine recording HSA failed. The data missing for these two subjects would be considered MCAR. However, the same mechanism is not operating for all missing values. For example, if we delete outliers, we will be doing so because they are too large. The mechanism generating those missing values will be missing not at random (MNAR, or nonignorably missing) by construction.

Another feature of the missingness in the gabapentin trial is that observations tend to be missing in chunks. In other words, if an observation at a particular time point is missing, it is more likely that the observation immediately following or preceeding it is missing than it is for an observation that is further removed in time. This pattern of missingness is called "wave nonresponse" in the survey sampling literature [4]. Most of the missing values in the gabapentin data are missing in "waves" rather than in well-separated points scattered throughout the subjects and time points. Furthermore, application of previous results in missing observation imputation for variance of estimators, consistency, bias, efficiency, etc, in the literature are asymptotic results.

Since the study is small, and we cannot afford to throw out missing values, we need a sensible way to replace those values. There are many ways to replace missing data. Some methods that have limited statistical validity include mean-imputation and last observation carried forward (LOCF). These have been shown to result in asymptotically biased estimates and decreased standard errors, which affect the inferences made from such data [3]. Regression mean imputation is a slightly more principled method of replacement [3], but it still results in deflated standard estimates of parameters.

4 TREATING THE MISSING DATA

Missing observations for the gabapentin data were treated in two stages. First, the last 24 hours of the HSA measurements were deleted. Only two subjects had complete 48-hour records for both pre- and post-treatment quantifications in this study. The original purpose in collecting 48 hours of observations was to test for a 24-hour rhythm to scratching activity. Even if those values had been replaced using a principled method, the resulting observations would probably not yield usable estimates of any such rhythm. Any estimates would have more nonresponse error than estimation error. Since most of the missing data occurred during these hours, deleting them decreased the percentage of missing data to 10% overall (not accounting for missing data caused by outlier deletion). Without the last 24 hours of measurements, the most important part of the study can still be salvaged: to make a decision about the effectiveness of gabapentin.

In the second stage, missing observations were replaced by an observation from a matching subject. This type of hot-deck imputation is sometimes called "Nearest Neighbor"

(NNHDI in the sequel) [5]. More precisely, let $y_i = (y_{i1}, \ldots, y_{ik})$ be a $K \times 1$ complete–data vector of outcomes. Further, let $y_i = (y_{\text{obs},i}, y_{\text{obs},m})$ where $y_{\text{obs},i}$ is the observed part and $y_{\text{obs},m}$ is the missing part of y_i. Then

$$\hat{y}_{it} = y_{\ell t} + (\overline{y}_{\text{obs},i} - \overline{y}_{\text{obs},\ell}) \tag{1}$$

where $\overline{y}_{\text{obs},i}$ is the mean of the observed values for subject i. Subject ℓ is the donor.

It is important to choose a donor that is "close" to the subject whose observations are missing. "Close" is defined by a metric, (*e. g.* $d(i,j) = \max_k |x_{ik} - x_{jk}|$) where $x_i = (x_{i1}, \ldots, x_{iK})^T$ are the values of K appropriately scaled covariates for a unit i at which y_i is missing [5]. For time series data, the distance metric is somewhat different, particularly since the donor choice will be made using a longitudinal variable.

Suppose subject i is missing a value at time t. For our purposes, the donor is defined as

$$d_j(t) = \min_j \sum_{t=1}^{T} |x_{it} - x_{jt}|, \tag{2}$$

for all $j = 1, \ldots, n-1$. Note that there are relatively few donors for the recipients in the gabapentin method. In our case, we do not use the same donor more than once. If one donor is chosen for two or more recipients, we use the next-nearest donor.

Donor subjects should be selected using another variable besides the variable which is being imputed. For the gabapentin study, visual analog scores (VAS), a measure of each subject's perception of scratching severity, were also measured every hour for 48 hours during the quantification period of the study. The VAS has been is used extensively in medical experiments as a way to measure outcomes such as pain and fatigue ([6, 7, 8]). Nearest neighbors were determined by computing the distance (2) between the recipient and all other subjects (candidate donors) on the basis of the VAS. Then, the values from HSA from the candidate donor with the minimum calculated distance was used to substitute for missing HSA observations in the recipient.

In typical hot-deck imputation, the missing observations are replaced with donated observations only once, and the new data are used as the "real" data set. This provides no estimate of imputation error, nor does it reflect the variability between subjects. For example, even if two subjects have exactly the same VAS trajectory, it is quite likely that their HSA values will be different due to random variation. It is necessary to estimate the uncertainty associated with replacement of missing values.

Two modifications were made to NNHDI. First, a random perturbation is added to mimic the inherent variability in the data. The random perturbation is generated from a $\mathcal{N}(0,29)$ distribution. The variance of the additive noise is the variance of the middle 80% of the extant HSA observations calculated over all subjects. In addition, three sets of imputed values are obtained, with three different sets of donors. Three sets provide enough information to estimate the imputation uncertainty, while keeping the analysis simple. The results of this analysis are given in Section 6. Before analysis can be attempted, it is important to deal with the issues of detection limits in the data.

5 UNDERSTANDING DETECTION LIMITS

In the previous section, the time-course plots of HSA revealed many hourly values which are identically zero. Recall that background levels were recorded for all subjects, then subtracted from the hourly scratching activity values. Some of those differences were negative.

The negative differences were recorded as zeroes. In some studies, replacing undesirable observations with zeroes would be inappropriate. However, it is important to consider what such a determination means for the data, and if it makes sense in context of the study. This type of thought is unique to the practice of statistics, and is typically learned through repeated exposure to the analysis of real data, rather than taught in any course.

The hourly HSA measurements do not measure scratching activity in the same way that a meter stick measures length. Indeed, the concept of "scratching activity" is not as well defined as the concept of length. HSA is the number of times that the frequency generated by the finger in the act of scratching crosses a certain threshold. It is proportional to the intensity of the scratch. However, there is no real zero; zero simply means that the amount of scratching during that hour was below the background level of body movement. One can even go so far as to say that HSA is really an interval-level variable. For any subject, the important thing about the measurement is the relative size of the pre-treatment value versus the post-treatment value. For this reason, the zeroes were not treated as detection limits, instead, they were left in the data as real values.

6 MIXED–EFFECTS MODEL ANALYSIS

The gabapentin experiment, with hourly scratching activity measurements collapsed into average pre and post measurements, can be seen as a split-plot design, where treatment (gabapentin or placebo) is the whole-plot factor, subjects are the whole plots, and the quantification time (baseline or post-treatment) are split-plot measurements. An equivalent analysis would be to consider this a repeated measures design, with baseline and post-treatment HSA scores being the repeated measures.

The model is given by

$$\mathbf{y_{ijk}} = \alpha_{\mathbf{i}} + \mathbf{b}_{j(i)} + \gamma_{\mathbf{k}} + (\alpha\gamma)_{ik} + \epsilon_{\mathbf{ijk}}, \tag{3}$$

where y_{ijk} is the response for the jth subject in the ith group at the kth quantification. The fixed effect, α_i, $i = 1, 2$, represents the effect of treatment group; b_j, $j = 1, 2$ is a random effect for the jth subject nested within the ith group, with $b_{j(i)} \sim NID(0, \sigma_b^2)$; γ_k is a fixed effect of the kth quantification, $k = 1, 2$; $(\alpha\gamma)_{ik}$ represents the fixed interaction effect between the ith treatment and the kth quantification, and $\epsilon_{ijk} \sim NID(0, \sigma^2 I)$.

Effect	DF	Mean Square	F-value	Pr > F
Group	1	8.38	16.57	0.0028
Subject (Group)	13	1.07	2.13	0.1290
Quant	1	6.07	12.00	0.0071
Group × Quant	1	6.42	12.70	0.0061

Table 1: ANOVA table with type III sums of squares for original (unmodified) data

For comparison purposes, the results for the analysis of the original, unmodified data are given in Table 1. In the table, the row labeled "Group" corresponds to the effect of the drug taken, and "Quant" corresponds to the time of quantification (pre-treatment or post-treatment). One can compute these results using SAS, Splus, R, or any other appropriate statistical software package.

Effect	DF	Mean Square	F Value	Pr > F
Group	1	1.74	8.29	0.0129
Subject (Group)	13	0.309	1.47	0.2480
Quant	1	0.469	2.23	0.1588
Group × Quant	1	1.32	6.31	0.0260

Table 2: Results for Average of 3 Imputations of NNHDI with additive $\mathcal{N}(0, 29)$ noise.

From this analysis, both the group effect, the quantification effect, and their interaction are highly significant. However, the analysis is what is called an analysis of the "completers" in this study. In other words, the data consist of those subjects for which complete measurements at both time points are available. Outliers are also present in these data, and may account for differences in the effects.

The results in Table 2 are those of model 3 applied to the average of the three data sets modified via NNHDI. These results account for missing values and outliers in the data. The outlying observations (those with large studentized error) were deleted from the data and replaced as if they were also missing observations, using the same donor as was used to replace the original missing values.

These results give different inferences than shown in Table 1. Here, the group effect and the interaction of group and quantification are still significant, but the quantification effect is no longer significant. This indicates that the outliers probably did have an effect on the inferences from the previous model.

The group and interaction effects in Table 2 are not as highly significant as those displayed in Table 1. This is likely due to nonresponse uncertainty. Little and Rubin (2002, pages 86-87) give a method for fraction of information about a parameter θ due to nonresponse (denoted γ). The larger the fraction, the more influence imputation has over the parameter estimates. It is applied here in order to obtain an idea of how much of the variability in the model can be attributed to the replacement of the missing values.

Let $\hat{\theta}_d$ and W_d, $d = 1, \ldots, D$, be D complete-data estimates and their associated variances for θ. $\hat{\gamma}_D = (1+1/D)B_D/T_D$ is an estimate of the fraction of information about θ due to nonresponse, where W_D is the within-imputation variance, B_D is the between-imputation variance, and T_D is the total variability across imputations. In the case of the gabapentin analysis, $D = 3$, and $\hat{\gamma}_D$ was less than one percent for the estimates of LME coefficients for group, quantification, and the interaction term. For the random effect of subject within group, approximately 52% of the information is due to nonresponse. This implies that the inferences we make from the imputed data for the fixed effects can be trusted. As for the random effect, more data and further analyses are needed before its importance can be ascertained.

7 SIMULATIONS FOR POWER AND SIZE

With a small data set, even with a small fraction of information due to missing observations, we still need to concern ourselves with power and size. The following simulations show how badly missing values can affect the power and size of a test. Thus, it is important to use some type of imputation. The simulations also show that there is a limit to the amount of data that can be imputed before inferences cannot be trusted.

We give simulations for two cases.

- Case 1: A Pretest/Posttest study with one normally distributed random variable ($\sigma^2 = 1$) and wave nonresponse in the data.
- Case 2: Wave nonresponse for longitudinal data with no correlation, analyzed with model 3.

For both cases, the size and power for 10%, 30%, and 50% missing values were compared, where the number of subjects was either 10 or 30. The power of each test (where the significance level was 0.05) was estimated under two effect sizes: a difference of two standard deviations between pre and post means, and a difference of five standard deviations. For all scenarios, the simulated examined a difference only in pre and post means, not between groups. Missing values were only in the post-treatment data, and they were replaced using the pre-treatment value plus standard Gaussian noise. Five-hundred replications of length 1000 were computed for each case.

	$N = 10$		$N = 30$		
% Missing	30%	50%	10%	30%	50%
$\mu_d = 0$	0.052	0.053	0.050	0.050	0.051
$\mu_d = 2$	0.662	0.341	0.999	0.995	0.904
$\mu_d = 5$	0.996	0.766	1	1	1

Table 3: Case 1: Paired t-test with Wave Nonresponse.

Table 3 displays these results. We see that power is quite poor when there are only 10 subjects. Having three or five observations missing in chunks is quite different from having three or five missing observations scattered throughout the data. Even when the difference in means is very large, there is a large decrease in power between 30% and 50% missing data for $n = 10$. It is not surprising that the t-test would have poor power when half of the data are missing. However, recall that the missing values have been replaced with randomly perturbed values. This scenario indicates that there is a limit to the prudent use of imputation techniques. It is interesting that the paired-test would do so well when 30% of the values are missing.

The power and size of a mixed–effects model (Model 3) from simulated data where we have longitudinal data (at 24 time points), but no correlation within subjects over time are given in Table 4. Note that a true two-standard deviation difference in pre and post means can be detected close to 100% of the time, even with 10 subjects and 50% missing observations replaced by imputed values. The size of the test for detecting pre and post differences is close to 0.05 for all combinations of missing observation percentages and sample sizes, as well.

	$N = 10$		$N = 30$	
Scenario	30%	50%	30%	50%
$\mu_d = 0$	0.053	0.049	0.051	0.035
$\mu_d = 2$	1	1	1	1

Table 4: Case 2: Size and Power for Longitudinal Data analyzed via Model 3.

8 DISCUSSION

Statistical practice is never as neat as textbooks sometimes imply. The challenge for a statistician is not only to analyze the data in a reasonable manner, but to obtain reasonable data to analyze.

This paper described how to prepare data from a small clinical trial where the data suffer from intermittent nonresponse, complete subject non response, and large outliers. For this trial, observations were missing through a mixture of mechanisms, which made modeling the nonresponse very difficult. As a simple and reasonable alternative, missing observations were replaced using a modified nearest-neighbor hot deck imputation (NNHDI). Outliers were also deleted, and subsequently replaced in the same manner. The resulting mixed–effects analysis produced reliable parameter estimates.

Simulations showed that the size and power are not unreasonable for sample sizes as small as 10 and percentage of missing observations less than 30%. In some cases, the power and size are very good, even when 50% of the data are missing and replaced by NNHDI.

References

[1] Bergasa NV, McGee M, Ginsburg I, and Engler D. Gabapentin treatment for the pruritis of cholestasis: results of a double-blind placebo-controlled trial. Poster presentation at the meeting of the American Association for the Study of Liver Disease: Boston, MA; November, 2004.

[2] Talbot TL, Schmitt JM, Bergasa NV, Jones EA, and Walker EC. Application of piezo film technology for the quantitative assessment of pruritis. Biomedical Instrumentation and Technology 1991; 25(5):400-403.

[3] Carpenter J and Kenward M. Economic and social research council missing data website. http:www.missingdata.org.uk. Date of Access: May 17, 2005.

[4] Pfeffermann D and Nathan G. Imputation for wave nonresponse: existing methods and a time series approach, in *Survey Nonresponse* (Groves, R. M., Dilman, D. A., Eltinge, J. L., and Little, R. J. A., eds.). New York: Wiley, 2002; pp. 417-430.

[5] Little RJA and Rubin DB. *Statistical analysis with missing data, 2nd ed.*. New York: Wiley Interscience, 2002.

[6] Aubrun F, Langeron O, Quesnel C, Coriat P, and Riou B. Relationships between measurement of pain using visual analog score and morphine requirements during postoperative intravenous morphine titration. *Anesthesiology*, 2003; **98**(6):1415-21.

[7] Hartmannsgruber MWB and Silverman DG. Applying parametric tests to visual analog scores. *Anesthesia & Analgesia*, 2000; **91**(1):248 - 249.

[8] Wewers ME and Lowe NK. A critical review of visual analog scales in the measurement of clinical phenomena. *Research in Nursing and Health*, 1990; **13**:227-236.