STATISTICS IN MEDICINE
Statist. Med. **17**, 1261–1291 (1998)

TUTORIAL IN BIOSTATISTICS

EXTENDING THE SIMPLE LINEAR REGRESSION MODEL TO ACCOUNT FOR CORRELATED RESPONSES: AN INTRODUCTION TO GENERALIZED ESTIMATING EQUATIONS AND MULTI-LEVEL MIXED MODELLING

PAUL BURTON[1,4]*, LYLE GURRIN[2,4], AND PETER SLY[3,4]

[1]*Division of Biostatistics and Genetic Epidemiology, TVW Telethon Institute for Child Health Research, P.O. Box 855, West Perth, WA 6872, Australia*
[2]*Princess Margaret Hospital for Children & King Edward Memorial Hospital, GPO Box D184, Perth, WA 6001, Australia*
[3]*Division of Clinical Science, TVW Telethon Institute for Child Health Research, P.O. Box 855, West Perth, WA 6872, Australia*
[4]*Department of Paediatrics, University of Western Australia, Nedlands, WA 6009, Australia*

SUMMARY

Much of the research in epidemiology and clinical science is based upon longitudinal designs which involve repeated measurements of a variable of interest in each of a series of individuals. Such designs can be very powerful, both statistically and scientifically, because they enable one to study changes *within* individual subjects over time or under varied conditions. However, this power arises because the repeated measurements tend to be correlated with one another, and this must be taken into proper account at the time of analysis or misleading conclusions may result. Recent advances in statistical theory and in software development mean that studies based upon such designs can now be analysed more easily, in a valid yet flexible manner, using a variety of approaches which include the use of *generalized estimating equations*, and *mixed* models which incorporate *random effects*. This paper provides a particularly simple illustration of the use of these two approaches, taking as a practical example the analysis of a study which examined the response of portable peak expiratory flow meters to changes in true peak expiratory flow in 12 children with asthma. The paper takes the reader through the relevant practicalities of model fitting, interpretation and criticism and demonstrates that, in a simple case such as this, analyses based upon these model-based approaches produce reassuringly similar inferences to standard analyses based upon more conventional methods. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Much of the research in epidemiology and clinical science is based upon longitudinal designs which involve repeated measurements of a variable of interest in each of a series of individuals. In

* Correspondence to: Professor Paul Burton, Division of Biostatistics and Genetic Epidemiology, TVW Telethon Institute for Child Health Research, P.O. Box 855, West Perth, WA 6872, Australia. E-mail: paulb@ichr.uwa.edu.au

Contract/grant sponsor: National Health and Medical Research Council of Australia

CCC 0277–6715/98/111261–31$17.50

the clinical or laboratory setting the measurements may take place over a relatively short period of time while longitudinal studies in epidemiology may span many years. Repeated measures designs can be very powerful, both statistically and scientifically, because they enable one to study changes *within* individuals over time or under a variety of different conditions.

Most standard statistical techniques, for example, the unpaired *t*-test, simple linear regression or the chi-square test for association, assume that each of the primary observations that make up a data set is independent of all of the others.[1,2] Unfortunately, this assumption can be inappropriate if repeated observations are taken within subjects[1–6] because observations within an individual tend to be correlated with one another. If one takes two observations at random from the same individual they are likely to be more similar in value than two random observations from two different individuals.[3,5–8] This means that each repeated observation in an individual may provide less additional information than a new observation in a new individual. This loss of information can be illustrated by the most extreme case; a measurement that varies between individuals but always takes exactly the same value in any one particular individual. In that case the *intra-class correlation* is 1·0 and each repeated observation conveys *no* additional information. Disregarding measurement error, height in middle-aged adults is an example of such a measure; if one wished to estimate mean height in this age group, one measurement in each of 100 subjects would be considerably more informative than ten measurements in each of ten.

It is clear that if a standard statistical analysis which assumes all observations to be independent is performed on repeated measures data when the intraclass correlation is positive, results may be misleading. For example, estimated standard errors are likely to be too small,[3,6] the analysis will effectively assume that there is more information in the data than there really is. Such an analysis has been referred to as *naive pooling*.[2,9]

Given that correlation can lead to a loss of information, it may seem surprising that repeated measures designs are used so commonly. However, when interest centres on a *change* in response under different conditions or over time, the longitudinal correlation between repeated observations means that within-person changes can be highly informative because they minimize the 'noise' arising from between-person variability. Thus, if one wished to test a new drug purporting to increase height in middle-aged adults, the fact that height is essentially constant in this age group means that the change in height within subjects (before drug versus after) will provide a powerful test of efficacy. In such circumstances, ignoring the correlation structure can waste important information and can make standard errors *too large*, as when an unpaired *t*-test is used on paired data with a positive intraclass correlation.[1]

Regardless of the direction of the error, it is obvious that a standard analysis which ignores an important correlation structure may well be misleading. Fortunately, valid approaches to analysis do exist. One approach is to resolve the repeated measurements in each individual to create a single *summary statistic*,[2] this approach, which may be referred to as *data resolution*, automatically avoids any over-inflation of the apparent size of the data set. The best statistic to use will depend upon the research question being asked. If the purpose of the repeated measurements is to average out within-person variability, the *mean* may be an appropriate statistic. On the other hand, if the repeated measurements are being used to monitor the change in value of a parameter of interest over time, a simple *difference* between two repeated measurements (as in a standard paired *t*-test[1]) or the *gradient* (slope) of a simple linear regression line[1] may be the appropriate summary statistics. This approach is safe and many statisticians recommend it when a research question may reasonably be addressed in such a manner.[2]

Unfortunately, many research questions cannot be answered with a single summary statistic. Furthermore, some summary statistics are inefficient; they use only part of the information in the data set. In such circumstances one requires a statistical technique that can extract the full information content of the data set without exaggerating the apparent sample size and yet can maintain the original structure of the data by avoiding the need to summarize the data *before* analysis. Fortunately, there are now a variety of ways to analyse repeated measures data in such a manner, that have been implemented in computer packages that are reasonably straightforward to use. These approaches include solutions based upon *generalized estimating equations*[7,8] (GEEs) and *multi-level modelling*[3,10] which is a form of linear mixed modelling,[10] the term mixed implying that the models incorporate *random effects* as well as conventional *fixed* regression coefficients.

Generalized estimating equations and mixed modelling are playing an increasingly important role in the analysis of studies in clinical science and epidemiology, and it is important that clinical researchers, epidemiologists and biological scientists are aware of their existence and that they have some idea how and when they might be used. This tutorial takes the reader through the practicalities of using generalized estimating equations and multi-level modelling to fit models for a particularly simple type of correlated data problem. The problem to be considered arises when repeated measurements are available on two continuous variables (one a response and one an explanatory variable) in each of a number of individuals (for example, people or animals) and primary interest centres upon the gradient of the association between the two variables. For example, (two real consulting problems) one may be interested in: (i) whether the slope of the relationship between pulse rate and blood pressure is flatter or steeper in experimental animals with impaired renal blood flow compared to controls; and (ii) whether the slope representing the increase in levels of a particular hormone over time is flatter or steeper in women who are destined to have a poor outcome to their pregnancy. Medical statisticians are commonly faced by research questions such as these. If the response variable is Normally distributed (or can be transformed to approximate Normality) the appropriate analysis is usually a generalization of simple linear regression, the generalization being required to address the non-independence, and resultant correlation, of the repeated measurements within each individual.

Taking a real example from our own research (Section 2), we will provide full details of the necessary data preparation and the practical model fitting procedures and will compare the results obtained using generalized estimating equations and multi-level modelling with the results of more conventional approaches. We supply computer code for statistical packages S-plus, SAS and mLn. This tutorial is intended as a practical guide for those who have not worked with generalized estimating equations and multi-level modelling before. We have deliberately kept mathematical notation as simple as possible and have avoided introducing matrix formulation. Readers who wish to properly understand the underlying theory at a more fundamental level should consider reading references 3–8 and 10.

2. A PRACTICAL EXAMPLE: ANALYSING THE RESPONSE OF PORTABLE PEAK FLOW METERS TO CHANGES IN PEAK EXPIRATORY FLOW

2.1. Background

Peak expiratory flow (PEF, the maximum rate of air flow in litres per minute during a forced expiration) is considered to provide an important measure of airway function. In particular, a fall

in PEF may provide warning of a forthcoming asthma attack while a patient is still asymptomatic. In recognition of this, a number of hand held PEF meters have been developed. Apart from a non-linear bias in the scale,[11–13] for which appropriate adjustment[13] can be made (see below), these have been shown to be both accurate and precise in the laboratory setting.[11–13] Accordingly, it has been recommended that asthmatic children should regularly monitor their PEF using such a meter.[11,12,14–16] However, there has been little work to investigate the accuracy, precision and clinical responsiveness of portable PEF meters when they are used by real children in the community.

Accordingly, we recently published[17] the results of a longitudinal study which investigated the response of portable peak expiratory flow meters to clinically relevant changes in true PEF identified using a Welsh–Allen (PneumoCheck) pneumotachograph spirometer as a gold standard measure for PEF. The study design was simple. Over a three month period in 1992, 12 asthmatic boys attending a boarding school in Western Australia made twice daily (pre-bronchodilator) estimates of PEF. Each measurement, which was supervised by a trained school nurse, consisted of a spirometric PEF estimate which was obtained using the pneumotachograph spirometer and a portable meter based estimate obtained using one of four different portable meters: Mini-Wright; Ferraris; Vitalograph; or Breath Taker.

The published analysis was simple and qualitative.[17] Our results demonstrated that the response of all PEF meters to clinically relevant falls in true PEF was remarkably poor. We concluded that the results of portable PEF monitoring should be treated with some circumspection and that further thought was required regarding the recommendation that portable PEF meters should form an integral part of the individual care plans for all children with asthma.[11,12,14–16]

Because the inferences that could be drawn from the qualitative analysis were necessarily limited, it was considered appropriate to proceed to a more formal quantitative analysis. This analysis had to take account of the longitudinal design of the study which meant that the results in each individual child were likely to be correlated with one another. Generalized estimating equations and multi-level modelling offered two possible approaches to the conduct of the analysis.

2.2. Defining the Statistics of Interest

A portable PEF meter may be poorly responsive to clinically relevant changes in true PEF for two reasons. First, a systematic fall (or rise) in observed meter PEF may be small relative to the fall (or rise) in true PEF that caused it. This represents an attenuated response and could occur, for example, if children were to adopt – as they may well do – habitual but sub-optimal peak flow techniques that do not properly reflect true PEF. Secondly, a poor clinical response might arise from excessive random within-child variation of meter PEF during periods of stable true PEF; excessive variation of this type might obscure the change in the meter readings when a fall in true PEF actually occurs.

If one is prepared to make the crucial assumption[18] that a pneumotachograph is an accurate and precise gold standard for true PEF, the scatter plot of a series of PEF measurements obtained using a portable meter which is also accurate and precise (vertical axis) versus a simultaneous series obtained using the pneumotachograph spirometer (horizontal axis) should produce a straight line of gradient one, passing through the origin. A gradient of less than one would suggest the meter to have an attenuated response. A vertical shift in the line would indicate a fixed additive bias. Any curvature would indicate a non-linear bias.

These considerations suggest that the statistics of interest should be: (i) the regression gradient of the relationship between meter and spirometer PEF; and (ii) the magnitude of random within-child variation. The magnitude of the correlation of results over time within a child is also of relevance although it is really a nuisance for which adjustment must be made rather than a parameter of primary interest. Note that a consistent additive bias need not necessarily impair the clinical value of a PEF meter[17] provided that it is still responsive to clinically relevant changes in peak flow.

It must be emphasized that if the assumption that the spirometer is a good gold standard is invalid, the stated approach to analysis would be seriously flawed.[18,19] In particular, if the spirometer readings were subject to serious random measurement error, one would anticipate regression to the mean and the estimated regression coefficients would be *expected* to be less than one.[19] However, the spirometer that was used met American Thoracic Society (ATS) criteria for accuracy and precision[12,13] and it was calibrated and used according to ATS guidelines. All PEF measurements were supervised by a trained nurse and readings that were registered as technically inadequate by the spirometer were rejected and repeated. Under such conditions, it is common, both in clinical practice and in research, to treat a pneumotachograph spirometer as a gold standard for PEF. Furthermore, a small repeatability study was undertaken and the empirical evidence indicated that the magnitude of the regression to the mean which might be anticipated in this setting was relatively small. Indeed, if the estimated regression coefficients we report in this paper were to be adjusted for this magnitude of bias, none of the substantive conclusions of the analysis would be changed. In order to avoid introducing a technicality which is peripheral to the main thrust of this paper we have therefore chosen to ignore this particular bias in our description of the analysis that was performed.

2.3. Data Preparation

For the sake of brevity, we restrict consideration to the analysis of the response of the Mini-Wright meter. Initial study confirmed the type of non-linear bias described previously to be a feature of many portable meters, including the Mini-Wright, with over-reading at low and moderate PEF and under-reading at high PEF.[11–13] Prior to analysis empirical correction algorithms were generated using the data reported by Miller *et al.*[12] and the observed meter readings were appropriately adjusted to take account of this non-linear bias. In the laboratory setting, it has been shown that having adjusted for this bias, portable PEF meters can fulfil ATS guidelines for measurement accuracy:[13] that is, $+/-12$ l/min or $+/-5$ per cent of true flow, whichever is the larger.[12] For the sake of clarity, the adjusted PEF measurements will be treated as if they were the raw data for the purposes of the analysis we report.

Before the commencement of modelling, the PEF measurements obtained with the spirometer and with the portable meter were approximately centralized [1] by subtracting 300l/min. Although centralization is good practice in any regression analysis,[1] it is our experience that failure to centralize leads to more serious problems (including total model fitting failure) in analyses based upon generalized estimating equations and random-effects models than it does in conventional regression analyses.

2.4. Statistical Analysis

2.4.1. Notation

Denote as **m** the variable holding the adjusted centralized portable meter PEF readings and as **s** the variable holding the centralized spirometer readings. Define m_{ij} as the jth adjusted

centralized portable meter reading in the ith child and s_{ij} as the corresponding spirometer reading in the same child.

2.4.2. *The correlation structure*

The analysis will investigate three possible structures for the correlation between the repeated observations in each child. First, the correlation will be ignored altogether and a *naively pooled* analysis will be carried out. Secondly, it will be assumed that the fundamental mechanism generating the correlation of observations within an individual is that, for any given true (spirometric) PEF, some children may tend to consistently record an overestimated portable meter PEF while others may tend to consistently record an underestimate. In other words, it will be assumed that the portable meter PEF readings in each child are potentially subject to a systematic additive bias (relative to true PEF), the magnitude of which varies between children. This correlation structure, which may be viewed as arising from a *varying regression intercept* will be modelled in three ways: (i) it will be modelled explicitly using *generalized estimating equations*; (ii) the variation between children will be modelled using a conventional *fixed-effects* model allowing a different intercept in each child – an analysis of covariance;[1] (iii) the variation between children in the systematic additive bias will be modelled using *random effects* in a *mixed* model – *multi-level modelling*. Thirdly, the correlation structure will be generalized by allowing the gradients as well as the intercept to vary between children and by allowing the variance of individual observations to depend upon true PEF. This analysis will be carried out using multi-level modelling and the resultant parameter estimates will be compared to those arising from a conventional analysis based upon *data resolution.*

2.4.3. *A naive pooled analysis*

The analysis was based upon standard simple linear regression taking **m** as the response variable and **s** as a sole explanatory variable. In conventional format, the resultant simple linear regression relationship may be expressed as $\mathbf{m} = \alpha + \beta\mathbf{s} + \mathbf{e}$ which is equivalent to:

$$m_{ij} = \alpha + \beta s_{ij} + e_{ij}. \tag{1}$$

Here α is the intercept at $\mathbf{s} = 0$ (300 l/min), β is the gradient (the parameter of interest) and **e** represents a Normally distributed error term uncorrelated with **s** which takes the value e_{ij} for the jth observation in the ith child.

Alternatively, the same regression relationship may be expressed in a convenient format commonly used for generalized linear models:

$$E(m_{ij}) = \alpha + \beta s_{ij}; \text{ error} = \text{Normal} \tag{2}$$

where $E(m_{ij})$ denotes the expected value of the jth portable meter reading in the ith child. This generalized linear model is said to have an *identity link* because its linear predictor, $\alpha + \beta s_{ij}$, directly predicts the expectation of the response variable rather than a function of it.

2.4.4. *Correlation arising from a varying intercept*

2.4.4.1. Generalized estimating equations (GEEs). The following comments relate to standard[7,8,20] generalized estimating equation models which may be called GEE1[21,22] models. There are some important differences between these models and more recent approaches such as

GEE2 and GEE4 but these are beyond the scope of this paper.[21,22] At the present time most GEE implementations in standard packages are GEE1. Our GEE1 modelling was carried out in *S-plus*[23,24] using the *gee*( ) function[25] which we obtained via e-mail from the *Statlib* archive in the U.K. The *gee*( ) function requires *S-plus* to interact with an external program written in C. We have used it exclusively in our UNIX-based implementation of *S-plus* and readers should be aware that one needs an appropriate C compiler to use it.

Generalized estimating equations extend generalized linear models,[26,27] which include simple linear regression, in two important ways. First, given a data set consisting of repeated measures, a GEE model allows the correlation of outcomes within an individual to be estimated and taken into appropriate account in the formulae which generate the regression coefficients and their standard errors. Secondly, GEE models permit the calculation of *robust* estimates for the standard errors of the regression coefficients.[6] Provided the basic linear regression relationship is correct[20] and there is no correlation in the measured responses *between* individuals,[5,7] robust standard errors ensure consistent inferences from a GEE1 model even if the chosen correlation structure (see below) is incorrect or if the strength of the correlation between repeated observations varies somewhat from individual to individual.[6–8] Although conventional *model-based* standard errors are also produced by most standard implementations of the GEE model, they are only consistent if the specified correlation structure is correct. In consequence, robust standard errors, which are often larger, are usually preferred. Needless to say, despite the assurance of consistency, the cost of choosing an incorrect correlation structure can be a loss of efficiency.[7] Robust standard errors are derived from what is sometimes called the *sandwich estimator* of the covariance matrix of the regression coefficients.[5] This was described[7,8] by Liang and Zeger in 1986, an equivalent formulation having been described earlier in 1967 by Huber.[28]

A standard GEE model is what is known as a *marginal* model[6,29] (see Section 3.2) and the basic regression relationship, in our case $E(m_{ij}) = \alpha + \beta s_{ij}$, and the within-subject correlation parameters are modelled completely separately[6] (orthogonally[21]). Accordingly, the process of fitting a GEE1 model may conveniently be viewed as a series of steps:[7]

(i) Fit a standard (naive) regression model assuming all observations to be independent.
(ii) Take the residuals from the regression and use these to estimate the parameters which quantify the correlation between observations in the same individual.
(iii) Refit the regression model using a modified algorithm incorporating a matrix which reflects the magnitude of the correlation estimated in step (ii).
(iv) Keep alternating between steps (ii) and (iii) until the estimates all stabilize.

Thus, in the simple linear regression case one might fit the model

$$E(m_{ij}) = \alpha^{[1]} + \beta^{[1]} s_{ij}; \text{ error} = \text{Normal}$$

where $\alpha^{[1]}$ and $\beta^{[1]}$ are the regression coefficients of the naive regression model, and then calculate the residuals for the jth observation in the ith child ($r_{ij}^{[1]}$) as

$$r_{ij}^{[1]} = m_{ij} - (\alpha^{[1]} + \beta^{[1]} s_{ij}).$$

These residuals may then be used to estimate the parameters which characterize the correlation between observations in an individual.[7] This information is then expressed in matrix form and incorporated into the estimating equations[7] which are used to generate new values for the regression coefficients $\alpha^{[2]}$ and $\beta^{[2]}$ and ultimately new residuals, $r_{ij}^{[2]}$, which are in turn used to

re-estimate the correlation parameters. The cyclical process continues until the estimates stabilize and convergence is achieved; this is guaranteed assuming standard regularity conditions.

Although the use of robust standard errors ensures that regression inferences are consistent regardless which correlation structure is chosen, there is no straightforward way in a GEE1 model to determine which is the *best* correlation structure to use. Consequently, one should choose the basic correlation structure to be used in the model *before* starting the analysis; one of the aims of model fitting is to estimate the quantitative parameter(s) which characterize the chosen structure. The *Appendix* lists some of the correlation structures which are commonly available in packages which implement GEE1 and indicates which quantitative parameters characterize each structure. The correlation structure chosen for our GEE analysis was *exchangeable*[7,8] which is also known as *compound symmetry*.[4] This means that every observation in an individual is assumed to be equally correlated with every other observation in that individual. This structure is fully characterized by one correlation parameter which is the intraclass correlation coefficient (see Appendix). If the correlation between observations in a subject *does* arise as a direct consequence of variation in the regression intercept between individuals, one would anticipate that all observations in an individual *would* be equally correlated and an exchangeable structure would therefore seem reasonable.

2.4.4.2. Analysis of covariance. As an alternative to explicitly specifying a correlation structure, the regression intercept can be permitted to vary from child to child by extending the regression model adopted in the naive analysis (2) to include a separate intercept for each child:

$$E(m_{ij}) = \alpha_i + \beta s_{ij}; \text{ error} = \text{Normal}. \tag{3}$$

Although this model allows a different value for α in each child it invokes a single value for β (a common gradient) and thus represents an analysis of covariance model which assumes parallelism.[1]

2.4.4.3. Multi-level modelling. In order to assist the reader, we start with notation which is similar to that used by Goldstein in Section 2.1 of the book *Multilevel Models in Educational and Social Research*;[3] the principal difference is that where Goldstein refers to σ^2 we use σ_e^2. Our later notation varies somewhat from Goldstein's book.[3]

Instead of specifying an exchangeable correlation structure explicitly (as in a GEE) or modelling the variation in intercept from child to child using conventional *fixed effects* (as in the analysis of covariance) one can instead extend a standard regression model by adding *random effects*.[3] In a standard regression model a regression coefficient is assumed to take the same fixed value for all individuals in a data set – hence the term 'fixed effect'. In contrast, random effects are regression coefficients that are permitted to vary from individual to individual. Thus, an alternative to permitting a separate fixed intercept in each child (as in the analysis of covariance) is to assume that there is some overall intercept (α_0) for the population of children in the study as a whole and that the discrepancy (u_i) between α_0 and the true intercept in the ith child (the 'effect of being child i') is generated from a Normal distribution with a mean of 0 and a variance of σ_u^2. The discrepancies (u_i) are called random effects: $u_i \sim \text{N}(0, \sigma_u^2)$.

The mixed (fixed and random effects) model which is equivalent to the analysis of covariance model (3) considered above may therefore be written as:

$$m_{ij} = \alpha_0 + \beta s_{ij} + u_i + e_{ij} \tag{4}$$

which, for convenience, is expressed in a notational form similar to (1) rather than (2). To fully characterize the model it should be noted that: $E(e_{ij}) = 0$; $\text{var}(e_{ij}) = \sigma_e^2$; and $\text{cov}(e_{ij}, e_{ik}) = 0$, where $E(e_{ij})$ and $\text{var}(e_{ij})$ denote the expectation and variance of the random error terms and $\text{cov}(e_{ij}, e_{ik})$ denotes the covariance between two error terms in the same individual. Similarly, $E(u_i) = 0$, $\text{var}(u_i) = \sigma_u^2$; and $\text{cov}(u_i, u_h) = 0$. Given this information and invoking basic theory,[3] the conditional variance of a single meter PEF reading is $(\sigma_e^2 + \sigma_u^2)$. Here the word 'conditional' indicates that we are referring to the variance of a meter reading (m_{ij}) about its expectation given the corresponding value of s_{ij} and the values of α_0 and β. Equivalently, the conditional covariance between two different meter readings in the same child is σ_u^2. Accordingly, the intraclass correlation coefficient may be estimated[3] as $\sigma_u^2/(\sigma_e^2 + \sigma_u^2)$. This means that by deriving estimates of σ_e^2 and σ_u^2 we are able to estimate, and therefore adjust for, the quantitative parameter that fully characterizes an exchangeable correlation structure (see Appendix).

One particularly convenient framework in which *mixed* models may be fitted is known as *multi-level modelling*,[3,10] an approach which has been implemented in the computer package *MLn*[30] which represents a generalization of *ML3*.[31] MLn is embedded in *Nanostat*[32] which was derived from *Minitab*.[33] Model fitting in MLn is based upon *iterative generalized least squares*.[3] As in the case of a GEE model, one may usefully consider the process of model fitting to consist of several steps:

(i) Fit a standard regression model assuming all observations to be independent (equivalent to assuming $\sigma_u^2 = 0$).
(ii) Take the residuals from the regression and use these to estimate σ_u^2 and σ_e^2.
(iii) Refit the regression model using a modified algorithm incorporating a covariance matrix which reflects the magnitude of σ_u^2 and σ_e^2 and hence takes account of the correlation structure.
(iv) Keep alternating between steps (ii) and (iii) until all estimates stabilize.

2.4.5. Extending the correlation structure, varying intercept and gradient

2.4.5.1. Multi-level modelling. The data structure arising from a study invoking repeated measures is usually hierarchical,[3] individual measurements (level 1) being made within subjects (level 2). Sometimes the measurements themselves form a hierarchy or the subjects may be clustered and in either case there may be a third or higher level. The judicious use of appropriate random effects at each level permits one to adjust for the influence of a wide variety of different correlation structures and to estimate variance, covariance and correlation terms that may be of specific interest in their own right.[3] In its most general form, the mixed model allows fixed and random regression coefficients to be associated with covariates observed at every level of a data hierarchy. The simplest of these models arises when there is a two level data hierarchy and, in addition to a random intercept, one wishes to allow the regression coefficient for a continuous covariate measured at level 1 to vary randomly between individuals (at level 2). This is precisely what we require to allow the gradient of the relationship between **m** and **s** to vary between children. The required model is a simple extension of model (4):

$$m_{ij} = \alpha_0 + (\beta_0 + v_i)s_{ij} + u_i + e_{ij} \tag{5}$$

where v_i is a level 2 random effect distributed $N(0, \sigma_v^2)$ which measures the discrepancy between the mean gradient in the population of children as a whole (β_0) and the true gradient in the ith child.

In order to allow the variance of individual observations to vary with s_{ij} we can also add an extra term at level 1:

$$m_{ij} = \alpha_0 + (\beta_0 + v_i)s_{ij} + u_i + e_{ij} + f_{ij}s_{ij} \tag{6}$$

where $f \sim \mathrm{N}(0, \sigma_f^2)$, if $\mathrm{cov}(e_{ij}, f_{ij}) = 0$ and the conditional variance of m_{ij}, around its predicted value given s_{ij}, α_0, β_0 and the level 2 random effects u_i and v_i, is $\sigma_e^2 + \sigma_f^2 s_{ij}^2$. This may be referred to as the *total level* 1 *variance*. The variance function may be generalized further by allowing a non-zero covariance term (σ_{ef}) between e_{ij} and f_{ij} and the *total level* 1 *variance* becomes $\sigma_e^2 + 2\sigma_{ef}s_{ij} + \sigma_f^2 s_{ij}^2$. We shall refer to this extended model as (6).

In a standard regression model, a *raw residual* is defined as the difference between the observed and predicted values of the dependent variable. There is only one error term, and, aside from their use in model checking, it is generally assumed that residuals arise as an uninteresting consequence of random variation. In a multi-level model, however, each random term defines a set of residuals and these may be of interest in their own right. Thus, the residuals associated with the random level 2 parameter which quantifies σ_v^2 estimate the amount by which the true gradient in each child differs from the overall estimated gradient[3,31] (that is, they estimate v_i) and may therefore be used to estimate the true gradient of the **m:s** relationship in each child individually (see Section 2.5.3).

The estimated residuals of a multi-level model are generated using a regression-based approach which takes each raw *total residual* (observed response minus the response predicted using the fixed parameters) and applies a series of *shrinkage factors*[3,31] which partition the total residual into components that may reasonably be attributed to each variance term at each level of the random structure of the model. With reference to the level 2 gradient residuals, this approach may conveniently be viewed as disentangling the proportion of each total residual that may be attributed to true variation in gradient from child to child, from that proportion which might better be attributed to random within-child variation or to other sources of variability at level 2. Because other sources of variation are taken into account, gradient estimates obtained using this method will, appropriately, have a tendency to be somewhat nearer the overall mean gradient than will estimates obtained from a series of simple linear regression analyses carried out in each child individually.[3] This may usefully be considered to reflect the fact that the children in whom the observed gradient is most extreme are likely to be those who not only have an extreme true gradient but in whom the random within-child variation takes the observed gradient even further in the same direction. On average, the observed gradient in such children will tend to be more extreme than the true gradient. In children with fewer observations the impact of within-child variability will, on average, be stronger and hence it is these children in whom the shrinkage factor tends to be greatest.[3] The method by which shrunken residuals are estimated assumes that the observed child-specific gradients are a random sample from a superpopulation of child-specific gradients which are Normally distributed with a mean of β_0 and a variance of σ_v^2. If there are good grounds for doubting the validity of this assumption,it would be unwise to use the shrunken residuals to estimate the child-specific gradients and it may then be preferable to estimate the gradient in each child individually, as in model (7) given below.

2.4.5.2. Data resolution. The conventional approach[2] to the analysis of these data would be to resolve the repeated measures in each individual by producing an appropriate summary statistic and then use that summary statistic as if it was the primary data. In this case, the appropriate summary statistic is the gradient of the simple linear regression line estimated in each child

individually. Thus, if β_k is the required gradient in the kth child, it is obtained as one component of the solution to the simple linear regression model fitted to that child's data alone:

$$E(m_{kj}) = \alpha_k + \beta_k s_{kj}; \text{ error} = \text{Normal}. \quad (7)$$

Having carried out an equivalent simple linear regression analysis in each child individually, the estimated gradient in each child ($\hat{\beta}_k$) may then be used as if it was the primary outcome variable. For example, one might describe the overall distribution of the $\hat{\beta}_k$ values or determine whether the mean of their distribution varies with an explanatory variable of interest.

To speed up estimation, the child-specific intercepts and gradients may all be obtained at the same time by extending the analysis of covariance model (3) to allow a separate gradient in each child. The resultant model structure would look the same as (7) but the model is fitted to all children simultaneously. Implicit in this extension of (3) is the assumption that the σ_e^2 are equal for all subjects. Fitting model (7) separately for each child permits a different σ_e^2 for each child.

2.4.6. Statistical inference

With the exception of GEE models, all tests of statistical significance associated with the addition (or deletion) of one or more fixed or random parameters to (or from) a regression model are based on the likelihood ratio test.[26,31] When drawing inferences about random parameters, it should be noted that to treat twice the change in log-likelihood between two nested models as asymptotically a χ^2 statistic can be misleading when the null hypothesis specifies that a single variance is exactly zero. This is because the null value then lies at the extreme of the range of possible values of a variance and this violates one of the standard regularity conditions needed to ensure that the asymptotic distribution of the test statistic is χ^2. In such cases, a standard χ^2 test will tend to understate the significance of an observed departure from the null hypothesis. Fortunately, all of the likelihood ratio test statistics for variances which are quoted in this paper are large and highly 'significant'. In consequence, despite this theoretical problem, we are confident that each of these null hypotheses *can* convincingly be rejected.

In the case of GEE models, formal inferences are based upon the Wald test; that is, having divided an estimated regression coefficient by its *robust* standard error, the result is treated as a *Normal standardized deviate* (Z^1). Conversely, the Z statistic may be squared and treated as a χ_1^2 statistic;[1] using the full variance-covariance matrix for the parameter estimates (not just the standard errors) this latter approach can be generalized to more than one regression coefficient.

2.5. Results

Although the results we report are restricted solely to the Mini-Wright meter, qualitatively equivalent results were found for all four meters.[17]

Over the period of study, 629 measurements that could be compared to a spirometer-based estimate of PEF were taken in 12 boys with the Mini-Wright meter. Table I details the relevant clinical characteristics of the study participants.

Figure 1 details the first and last five lines of the primary data file used in all of the analyses we report. All of the computer packages we used can read a plain ASCII text file in free format.

2.5.1. Naive pooled analysis

Figure 2 is a two-way scatter plot illustrating the observed relationship (naively pooled over subjects) between the Mini-Wright and spirometer based estimates of PEF; the estimated simple

Table I. Clinical characteristics

Child	Age (years)	Asthma severity*	Treatment†	Number of Mini-Wright observations paired with a spirometer reading
1	13	Moderate	S, B	101
2	15	Moderate	S	46
3	12	Mild	Nil	44
4	12	Mild	S	43
5	11	Moderate	S, T	44
6	15	Severe	S, B	59
7	12	Mild	S	28
8	17	Mild	S	33
9	11	Moderate	S,B	44
10	12	Labile	S	54
11	14	Mild	S,B	62
12	16	Severe	S,B,T	71

* Severity of asthma judged clinically
† Treatment: S = salbutamol; B = beclamethasone dipropionate; T = theophylline

Child	Observation	m	cons	s
1.0000	1.0000	134.01	1.0000	192.00
1.0000	2.0000	146.38	1.0000	204.00
1.0000	3.0000	109.85	1.0000	236.40
1.0000	4.0000	419.25	1.0000	226.80
1.0000	5.0000	197.74	1.0000	171.60
		. . . . .		
		. . . . .		
12.000	67.000	146.38	1.0000	63.000
12.000	68.000	86.410	1.0000	60.000
12.000	69.000	134.01	1.0000	149.40
12.000	70.000	98.040	1.0000	161.40
12.000	71.000	74.970	1.0000	149.40

Figure 1. Layout of primary data set: PEF.DAT

linear regression line, generated using model (2), is superimposed. Confidence limits for $E(m_{ij})$ at a given value of s_{ij} are also superimposed; they were obtained in the standard manner.[1] The estimated gradient of the simple linear regression line is 0·762 with standard error = 0·0394, generating an approximate 95 per cent confidence interval ($0{\cdot}762 \pm 1{\cdot}96 \times 0{\cdot}0394$) of 0·68 to 0·84. This result could naively be interpreted as suggesting that portable PEF metres *are* adequately responsive to changes in true PEF and that, to be specific, a 100 l/min fall in true PEF will, on average, be associated with a 76 l/min fall in meter based PEF and that, in most cases, the clinical message of the meter will therefore be correct.

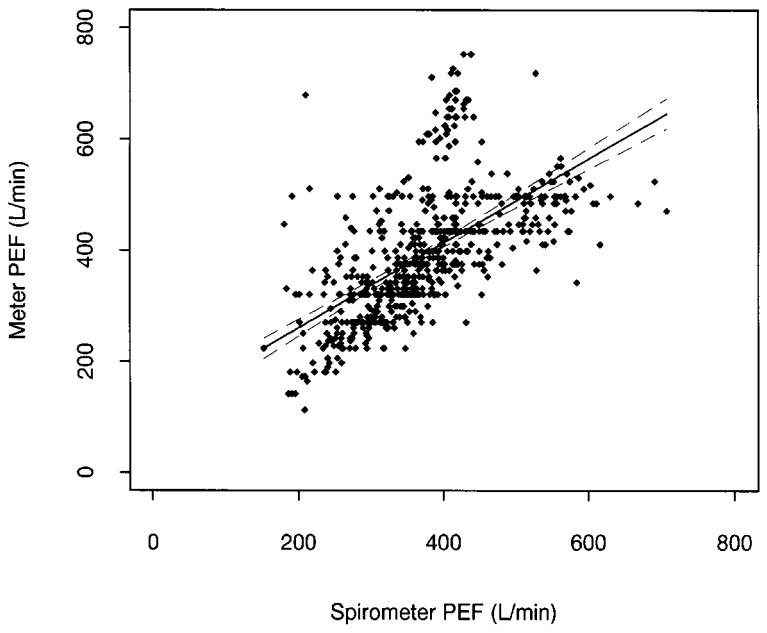

Figure 2. An analysis based upon naive pooling. A scatter plot of portable meter PEF versus spirometer PEF. The simple linear regression line (———) and the 95 per cent confidence interval for $E(m_{ij})$ at a given value of s_{ij} (– – – – –) are superimposed upon the raw data (◆)

2.5.2. *Correlation arising from a varying intercept*

2.5.2.1. Generalized estimating equations. The GEE approach was used to extend model (2). As would be expected, there was strong evidence ($Z = 3{\cdot}79$, $p = 0{\cdot}00015$) of a linear relationship between **m** and **s**, however, there was little evidence of quadratic ($Z = 1{\cdot}42$, $p = 0{\cdot}16$) or cubic ($Z = 0{\cdot}25$, $p = 0{\cdot}80$) curvature. This suggests that the adjustment for non-linear bias (see Section 2.3) was successful.

Figure 3 details the *S-plus* code which was used to read in the primary data file and to fit a GEE model with an exchangeable correlation structure (see Appendix). Table II details the estimated parameters of this model. The estimated intraclass correlation coefficient is 0·815. Having taken account of this correlation the estimated gradient of the relationship between **m** and **s** is considerably flatter (0·247 [0·120 to 0·370]) than the estimate from the naive analysis. This result indicates that a 100 l/min fall in true PEF will, on average, be associated with a 24·7 l/min fall in meter PEF. Figure 4 details the code required to fit the equivalent GEE model in *SAS*[34] using the *geel_203 macro*.[35] Parameter estimates are very similar to those reported in Table II, the largest discrepancy being in the estimate of the intraclass correlation coefficient 0·815 in S-plus, 0·825 in SAS. It should be noted that older implementations of GEE1 use unweighted averaging schemes for calculating correlation parameters and a minor difference might therefore be anticipated.[35]

2.5.2.2. Analysis of covariance. The standard *glm*( ) function in S-plus was used to fit model (3), an analysis of covariance model with a common gradient. The estimated common gradient ($\hat{\beta}$)


```
#       Read in data and put it into a data frame called 'pef'
pef <- scan("pef.dat",what=list(child=0,obs=0,m=0,cons=0,s=0))

#       Link Splus to compiled object code for C program which fits model
dyn.load.shared("/biostat/paulb/gee/cgee.so")

#       Fit model
gee.mod.exch  <-  gee(x=pef$s, y=pef$m, id=pef$child, varfun="gaussian", link="identity",
                      corstr="exchangeable", intercept=T, summarize=F)
#       Modelling output directed to an object called gee.mod.exch (an arbitrary name)
#       x = explanatory variable,
#       y = response variable,
#       id=cluster (individual) identifier,
#       varfun = distribution of response variable (Normal),
#       link = link function (identity),
#       corstr = correlation structure (exchangeable),
#       intercept=T   fits a regression intercept
#       summarize=F allows output to be saved as an Splus object (i.e. gee.mod.exch)

#       To interpret output:
#       Robust and model-based standard errors can only be obtained from the square
#       root of the diagonal elements of the robust and model-based variance-covariance
#       matrices called  "gee.mod.exch$robvar" and "gee.mod.exch$naivvar"
#       respectively. Regression coefficients are held in the vector "gee.mod.exch$regest".
#       The intra-class correlation coefficient may be obtained from the off-diagonal
#       elements of the working correlation matrix called "gee.mod.exch$worcor".
```

 # denotes a comment in *Splus*. Unbolded text represents active *Splus* code.

Figure 3. S-plus code required to fit an GEE model with an exchangeable correlation structure

Table II. A GEE model with an *exchangeable* correlation structure

Parameter	Value
Fixed regression coefficients	
α the overall regression intercept	59·2
(*Robust* standard error)	(27·1)
[95 per cent confidence interval]	[6·1 to 112·3]
{*Model-based* standard error}	{26·14}
β the gradient	0·247
(*Robust standard* error)	(0·065)
[95 per cent confidence interval]	[0·120 to 0·374]
{*Model-based* standard error}	{0·031}
Correlation coefficient	
Intraclass correlation	0·815


```
/*
        Read in data and put it into a SAS data set called 'pef'.
*/
data pef ;
infile 'pef.dat';
input child obs m cons s;
run;
/*
        Fit model using SAS gee1_203  macro. Macro gee1_203 is written using SAS
        IML.  Proc IML and SAS macro language required.
*/
%include  'c:\sas\gee\gee1_203.sas';
%gee(data=pef,  yvar=m,  xvar=cons s,  id=child, link=1, vari=1, corr=4);
/*
        data=data set name,
        y var= response variable,
        x var= explanatory variables (intercept must be fitted explicitly using a column of
                1s [cons])
        id=cluster (individual) identifier,
        link = link function (1=identity),
        vari = distribution of response variable (1=normal),
        corr = correlation structure (4=exchangeable).

        To interpret output:
        Regression coefficients printed out with robust and model based standard errors.
        Working correlation structure printed out as a matrix. For exchangeable
        correlation structure off-diagonal elements give intra-class correlation.
*/
```

/* denotes a comment in *SAS*.*/ Unbolded text represents active *SAS* code.

Figure 4. SAS code required to fit a GEE model with an exchangeable correlation structure

was 0·240 [0·175 to 0·305] which was very similar to its GEE counterpart (see Table II). Although the standard error (0·033) was rather smaller than its GEE equivalent this was in part a consequence of using *robust* standard errors in the GEE model. The equivalent *model-based* standard error in the GEE model (0.031) was very similar to the estimate from the analysis of covariance. The deviance (residual sum of squares[27]) about model (3) was 1,224,320 on 616 degrees of freedom. When the βs_{ij} term was deleted from the model, the deviance increased to 1,331,047 on 617 degrees of freedom. This represented a change of 106,727 and having divided this by the scale parameter derived from the fuller model ($1{,}224{,}320 \div 616 = 1987{\cdot}5$), it equated to a change in scaled deviance of $106{,}727 \div 1987{\cdot}5 = 53{\cdot}7$. This is equivalent to twice the change in log-likelihood between the two models and may be compared (approximately) to a χ^2_1 distribution.[26,27] This provides strong evidence ($p < 0{\cdot}00001$) of a linear relationship between **m** and **s**. The square of the Z statistic which provides a formal test for the equivalent relationship in the GEE model was $3{\cdot}79^2 = 14{\cdot}36$, but if one uses the *model-based* standard error rather than the *robust* estimate, the squared Z statistic is 63·5 which is not dissimilar to 53·7.

2.5.2.3. Multi-level modelling. Figures 5 to 7 shows the MLn code required to read in the primary data file and to carry out the main MLn analyses discussed in this paper.


```
NOTE [A]    Read in data and name columns.
dinp c1-c5
pef.dat
name c1 'child' c2 'obs' c3 'm' c4 'cons' c5 's'

NOTE [B]    Start up a log file to record session.
logo pef.log

NOTE [C]    Set up initial model structure: response variable, explanatories, level 2 and
NOTE        level 1 identifiers (variables coding individuals and observations) and
NOTE        variance terms at level 2 and level 1 for σu² and σe².
resp 'm'
expl 'cons' 's'
iden 2 'child'
iden 1 'obs'
setv 2 'cons'
setv 1 'cons'

NOTE [D]    Set up modelling environment: Run multiple iterations (batch) with maximum
NOTE        100 iterations. You can always break out of a model fit with the escape key.
batch
maxiter 100

NOTE [E]    Look at summary of model which has been set up
sett

NOTE [F]    Start model fit
start

NOTE [G]    Look at fixed and random parameter estimates and calculate likelihood.
NOTE        The command 'like' calculates minus twice the log-likelihood of the model.
fixe
rand
like
NOTE        This is model (M4)

NOTE [H]    Look at significance of linear term in 's'. The 'expl' command is a toggle.
NOTE        If a variable is already in the model it will remove it. So in this case
NOTE        the first 'expl' directive removes 's' the second restores 's to the model.
expl 's'
next
like
expl 's'
```

NOTE denotes a comment in *MLn*. Unbolded text represents active *MLn* code.

Figure 5. MLn code required to fit model (4) and to test significance of linear relationship between **m** and **s**

Steps [C]–[G] in Figure 5 set up and fit model (4). Using the 'LIKElihood' command MLn reports 'minus twice the log-likelihood' of model (4) as 6624·89 and the equivalent value for the corresponding model with the βs_{ij} term removed (step [H]) as 6679·88. The likelihood ratio test statistic is therefore $\chi^2_1 = 54{\cdot}99$, $p < 0{\cdot}00001$. This is very similar to the analogous result ($\chi^2_1 = 53{\cdot}7$) from the analysis of covariance. So, as with the GEE and analysis of covariance, multi-level modelling provides strong evidence of a linear relationship between **m** and **s**.


```
NOTE [I]  Extend model to allow a separate gradient in each child.
NOTE      Random terms can be added using 'setv'ariance or 'sete'lement. The command
NOTE      setv k 'x' adds all possible variance and covariance terms at level k which involve 'x'.
NOTE      The command sete  k 'x' 'x' adds only the variance term for 'x' while sete k 'x' 'y'
NOTE      adds only the covariance between 'x' and 'y'.
sete 2 's' 's'
next
like
NOTE      This is now model (M5).

NOTE [J]  Allow level 1 variance to depend on 's'. Look at change in likelihood.
sete 1 's' 's'
next
like
sete 1 's' 'cons'
next
like
NOTE      This is now 'final' model (M6)

NOTE [K] Look for need to extend model (M6) with covariance term at level 2.
sete 2 's' 'cons'
next
like

NOTE [L]  Clear covariance term from level 2. The command 'clre'lement  removes single term.
clre 2 's' 'cons'
next

NOTE [M] Look at change in likelihood if square and cubic terms in 's' are added to the model.
calc c6='s'**2
calc c7='s'**3
name c6 's_sq' c7 's_cub'
expl 's_sq'
next
like
expl 's_cub'
next
like

NOTE [N] Refit model (M6) without square and cubic terms.
expl 's_sq' 's_cub'
next
fixe
rand
```

NOTE denotes a comment in *MLn*. Unbolded text represents active *MLn* code.

Figure 6. MLn code required to fit models (5) and (6) and to test significance of adding additional covariance term at level 2 and non-linear terms in **s**

Table III reports the parameter estimates for model (4). The random parameters at level 1 and 2 estimate σ_e^2 and σ_u^2, respectively. Their joint inclusion in the model is equivalent to permitting an exchangeable correlation structure (see Appendix). As can be seen, the fixed regression coefficients are the same as their equivalents in the GEE model (see Table II) and the intraclass


```
NOTE [O]    Set up calculation of child-specific gradients: Generate residuals and adjusted comparative
NOTE        variances. For details see MLn command reference (version 1.0: march 95) page 43.
rlev 2
rtyp 2
rcov 1
rout c100-c103
resi

NOTE [P]    Calculate child-specific gradients, standard errors and 95% confidence intervals.
NOTE        Overall gradient=0.386, standard error=0.088.
calc c104=c101+0.386
calc c105=sqrt(c103+(0.088**2))
calc c106=c104-1.96*c105
calc c107=c104+1.96*c105
name c104 'GRADIENT' c105 'GRAD_SE' c106 'GRAD_L95' c107 'GRAD_U95'
print c104-c107

NOTE [Q]    Calculate absolute and percentage critical falls for table 5.
tabs c110 's' 'child'
calc c111=1.28*sqrt(926.8-2*2.156*c110+0.128*c110**2)/'gradient'
calc c110=c110+300
round c110 c110
round c111 c111
calc c112=100*c111/c110
round c112 c112
name c110 'mean_s' c111 'absfall' c112 'percfall'
print c110-c112

NOTE [R]    Generate diagnostically standardised residuals for level 1.
rlev 1
rtyp 0
rcov 1
rout c150-c153
resi
calc c154=c150/sqrt(c152)
calc c155=c151/sqrt(c153)
name c154 'sres_c1' c155 'sres_s1'

NOTE [S]    Generate fitted values for checking level 1 diagnostic residuals. Fitted value prediction
NOTE        uses fixed effects (for 'cons' and 's') and level 2 residuals (c100 and c101 [from step O]).
predict 'cons' c100 's' c101 c170

NOTE [T]    Level 1 diagnostic plots (histogram, normal plot, plot v fitted values).
hist c154
nscores c154 c156
plot c154 c156
plot c154 c170
NOTE        Do the same for c155 and then repeat steps [R] -> [T] for level 2 diagnostic residuals.
```

NOTE denotes a comment in *MLn*. Unbolded text represents active *MLn* code.

Figure 7. MLn code required to calculate child-specific gradients, standard errors and 95 per cent confidence intervals and to conduct diagnostic residual analysis

Table III. A multi-level model with an *exchangeable* correlation structure

Parameter	Value
Fixed regression coefficients	
CONS*	59·2
α_0 the overall regression intercept	(26·2)
(Standard error)	[7·8 to 110·6]
[95 per cent confidence interval]	
S*	0·247
β the common gradient	(0·033)
(Standard error)	[0·182 to 0·312]
[95 per cent confidence interval]	
Level 2 random components	
CONS/CONS*	8110
σ_u^2 variance of intercept between children	(3298)
(Standard error)	
Level 1 random components	
CONS/CONS*	1984
σ_e^2 variance within a child	(113)
(Standard error)	

* Names given to parameters by MLn

correlation coefficient may be estimated as $8110/(8110 + 1984) = 0{\cdot}803$ which is also very similar to the estimate (0·815) obtained from the GEE model. On the other hand, the estimated standard errors for the fixed regression parameters are smaller than their equivalents in the GEE model. This is again because *robust* standard errors are quoted for the GEE model and the MLn standard errors are very similar to their *model-based* counterparts in the GEE analysis (see Tables II and III).

2.5.3. *Extending the correlation structure*

2.5.3.1. Multi-level modelling. Step [I] in Figure 6 extends model (4) to produce model (5) which permits a different gradient in each child. The likelihood ratio test statistic associated with the extension is 20·97. Treated as a χ_1^2 this is very large and clearly significant, $p < 0{\cdot}00001$. This indicates that the gradient of the relationship between **m** and **s** varies randomly between subjects considerably more than would be expected by chance alone given the structure of model (4). That is, it suggests that some children have a truly steeper gradient than others.

Step [J] further extends model (5) by allowing the variance at level 1 to depend upon **s**. A significant improvement in model fit accompanies both the addition of a term $f_{ij}s_{ij}$ ($f_{ij} \sim \mathrm{N}(0, \sigma_f^2)$) to the model ($\chi_1^2 = 35{\cdot}02$) and the inclusion of a non-zero covariance term (σ_{ef}) between e_{ij} and f_{ij}($\chi_1^2 = 6{\cdot}41$, $p = 0{\cdot}01$). This is model (6). Inclusion of a non-zero covariance term (σ_{uv}) between u_i and v_i at level 2 (step [K]) does not significantly improve the fit of the model ($\chi_1^2 = 0{\cdot}0$, $p = 1$) and in step [L] model (6) is restored by clearing the unwanted covariance term using the 'CLRElement' command. Step [M] tests for a non-linear relationship between **m** and

Table IV. An extension of the multi-level model to permit random variation between subjects of the gradient between **m** and **s** and to account for the dependence of within-child variance upon **s**

Parameter	Value
Fixed regression coefficients	
CONS*	53·7
α_0 the overall regression intercept	(22·5)
(Standard error)	[9·6 to 97·8]
[95 per cent confidence interval]	
S*	0·386
β the overall gradient	(0·088)
(Standard error)	[0·214 to 0·558]
[95 per cent confidence interval]	
Level 2 random components	
CONS/CONS*	5771
σ_u^2 variance of intercept between children	(2442)
(Standard error)	
S/S*	0·0670
σ_v^2 variance of gradient between children	(0·0372)
(Standard error)	
Level 1 random components	
CONS/CONS*	926·8
σ_e^2 intercept component of level 1 variance	(93·5)
(Standard error)	
CONS/S*	−2·156
σ_{ef} intercept/gradient covariance at level 1	(1·12)
(Standard error)	
S/S*	0·128
σ_f^2 gradient component of level 1 variance	(0·019)
(Standard error)	

* Names given to parameters by MLn

s but, as in the GEE model, neither square ($\chi_1^2 = 0{\cdot}04$, $p = 0{\cdot}84$) nor cubic terms ($\chi_1^2 = 0{\cdot}53$, $p = 0{\cdot}47$) in **s** significantly improve the fit of the model. In step (N), model (6) is restored and is adopted as the final model from which inferences are to be drawn. Table IV details the parameter estimates of model (6).

2.5.3.2. Computing the statistics of principal interest. The random parameters at level 1 of model (6) jointly indicate that the within-child variability of **m** varies with **s**. Using the random parameters at level 1 (see Table IV) and invoking standard variance-covariance theory,[3] it may be estimated that given a stable spirometric PEF of $\overset{\circ}{V}$ l/min, the estimated standard deviation ($\widehat{\mathrm{SD}}$) of portable meter PEF readings within a child will be

$$\widehat{\mathrm{SD}} = \sqrt{\{926{\cdot}8 + 2 \times (-2{\cdot}156) \times (\overset{\circ}{V} - 300) + 0{\cdot}128 \times (\overset{\circ}{V} - 300)^2\}}. \quad (7)$$

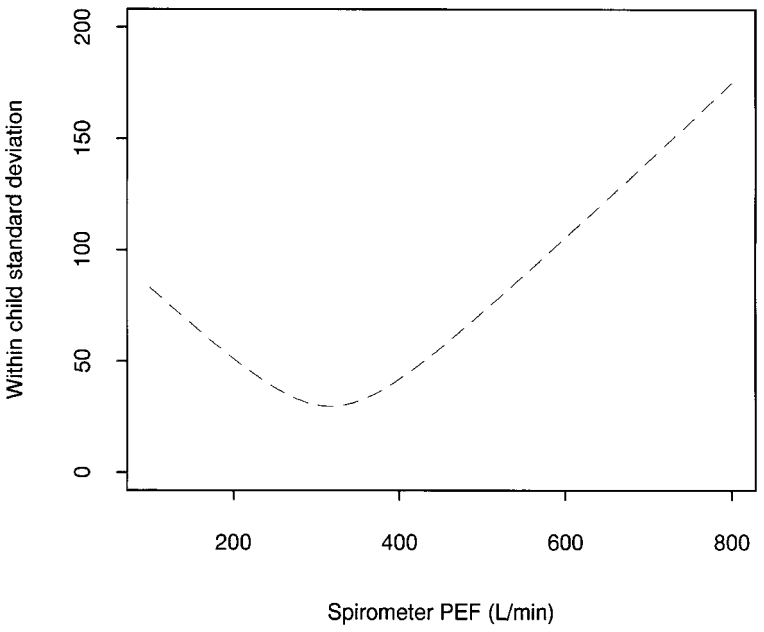

Figure 8. The estimated standard deviation of the random within-child variation of meter PEF (when spirometer PEF is stable) over a range of possible values of spirometer PEF

This indicates that, given a stable spirometric peak flow, the within-child variability of portable meter readings increases markedly at high values of **s** and to a lesser extent at low values (see Figure 8 for a graphical representation). Minimum variability is estimated to occur at $\overset{\circ}{V} = 316{\cdot}8$ l/min when $\widehat{\mathrm{SD}} = 29{\cdot}8$ l/min.

Given an estimated overall **m:s** gradient of $\hat{\beta}_0$ (that is, 0·386, Table IV) and given that the estimated level 2 residual for the **S**/**S** term in the ith child is $\hat{v}_i$, the true gradient in the ith child ($\hat{\beta}_i$) may be estimated[3,31] as $\hat{\beta}_i = \hat{\beta}_0 + \hat{v}_i$ and the estimated standard error of this gradient may be obtained as

$$\hat{\sigma}(\hat{\beta}_i) = \sqrt{\{\hat{\sigma}^2(\hat{\beta}_0) + \hat{\sigma}_c^2(\hat{v}_i)\}}$$

where $\hat{\sigma}^2(\hat{\beta}_0)$ is the squared estimated standard error of the fixed coefficient for **S** (that is, $0{\cdot}088^2 = 0{\cdot}0077$, Table IV) and $\hat{\sigma}_c^2(\hat{v}_i)$ is the *comparative* variance of the relevant residual in the ith child (Goldstein, 1987, page 48).[3] *Comparative* variance for residuals may be obtained directly in MLn[30] and by selecting 'RTYPe2' (see Figure 7 step [O]) one obtains comparative variances which have been adjusted for the fact that the random parameters from which they are computed are themselves estimates.[3,30] An approximate 95 per cent confidence interval (see Table V) for the estimated gradient in the ith child may then be obtained as

$$\hat{\beta}_i \pm 1{\cdot}96 \times \hat{\sigma}(\hat{\beta}_i).$$

Figure 7 steps [O] and [P] detail the MLn code required to generate child specific gradients, standard errors and 95 per cent confidence intervals. Figure 9 plots the raw data for each child:

Table V. The estimated gradient [and 95 per cent confidence interval] for the relationship between **m** and **s** in each individual child, and the estimated magnitude of fall in true PEF that might reasonably be detected with the portable meter despite random within-child variability

Child	Estimated gradient [95% confidence interval]	Mean spirometric PEF (l/min)	Absolute detectable fall (l/min)	Percentage detectable fall
1	0·15 [−0·12 to 0·41]	437	453	104%
2	0·40 [0·05 to 0·76]	312	95	30%
3	0·60 [0·27 to 0·94]	269	73	27%
4	0·22 [−0·17 to 0·60]	352	191	54%
5	0·55 [0·20 to 0·91]	252	87	35%
6	0·57 [0·17 to 0·98]	391	89	23%
7	0·52 [0·16 to 0·89]	324	73	23%
8	0·34 [−0·06 to 0·75]	563	346	61%
9	0·13 [−0·19 to 0·44]	317	304	96%
10	0·10 [−0·21 to 0·42]	332	373	112%
11	0·82 [0·50 to 1·14]	396	64	16%
12	0·22 [−0·11 to 0·55]	422	284	67%

the estimated overall regression relationship $E(m_{ij}) = \hat{\alpha}_0 + \hat{\beta}_0 s_{ij}$; the estimated child-specific regression relationship in the kth child $E(m_{kj}) = \hat{\alpha}_0 + \hat{u}_k + (\hat{\beta}_0 + \hat{v}_k)s_{kj}$; and the modelled 2·5 per cent and 97·5 per cent percentiles for the observed meter readings in the kth child calculated as

$$E(m_{kj}) \pm 1{\cdot}96 \times \sqrt{(\hat{\sigma}_e^2 + 2\hat{\sigma}_{ef} s_{kj} + \hat{\sigma}_f^2 s_{kj}^2)}. \tag{8}$$

2.5.3.3. Interpreting the model. Having used the multi-level model to derive values for the statistics of principal interest there are a number of approaches which one might take to the combination of information about the magnitude of random within-child variation and the magnitude of the **m:s** regression gradient in order to draw conclusions about the potential utility of portable PEF monitoring. We will restrict our attention to one possible approach. Although we believe this method to be sensible, we do not claim that it is 'optimal' in any technical sense of the word.

In order for a child to respond appropriately to a reduction in meter PEF following a period of stable true (spirometric) PEF, it is necessary for the observed change to be discriminated from the background noise representing random within-child variation of **m** around its expected value during the period of stability. On the basis of an empirical review of a series of computer

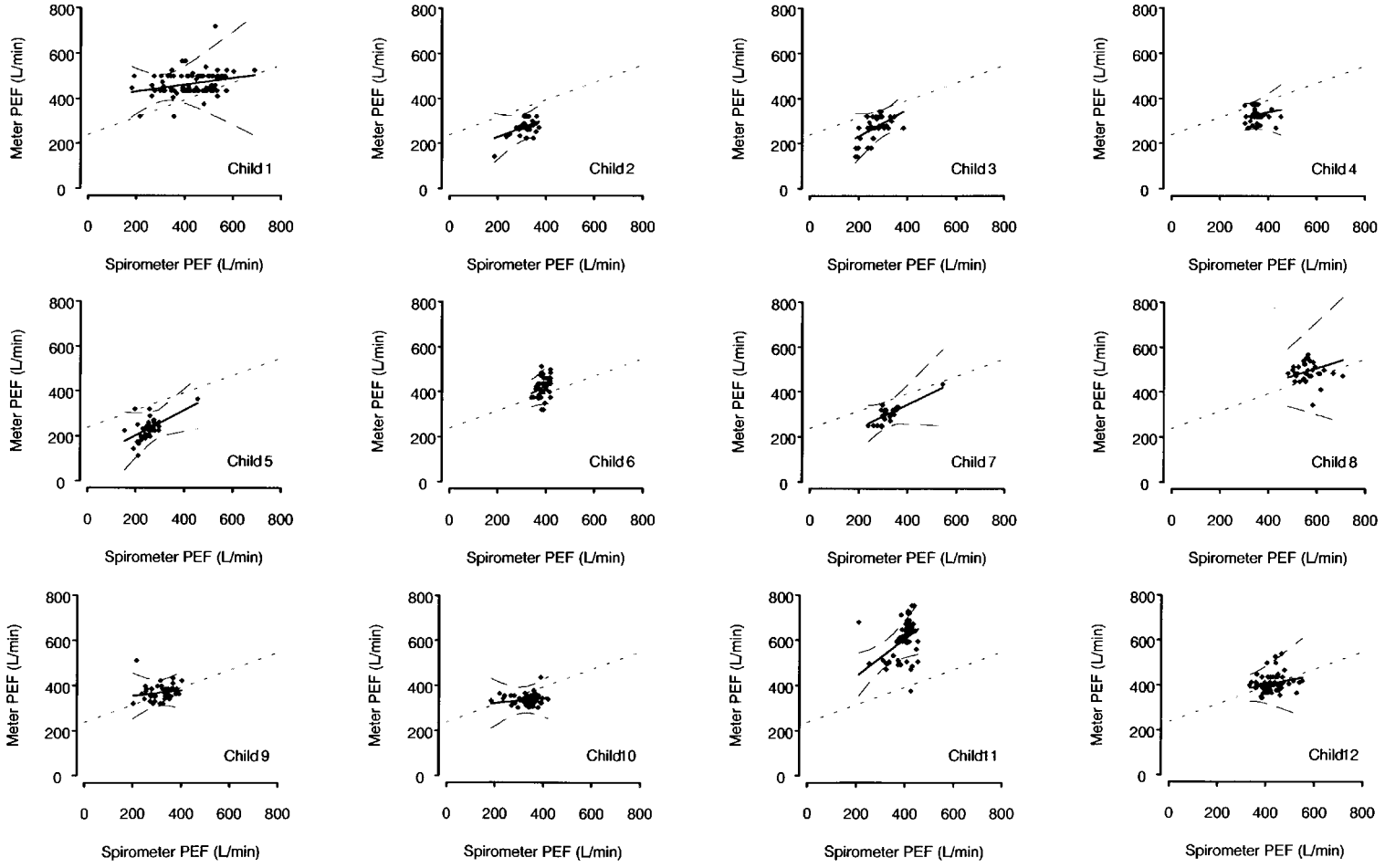

Figure 9. A plot of portable meter PEF versus spirometer PEF in each child individually. raw data (◆); estimated overall regression relationship (- - - - -); estimated child-specific regression relationship (———); modelled 2·5 per cent and 97·5 per cent percentiles for the observed meter readings (– – – – –)

simulations, we believe that following a period during which **s** is stable and the estimated within-child standard deviation of **m** is $\widehat{\mathrm{SD}}$, the estimated minimum fall in the expectation of **m** which has a reasonable chance of being detected within two PEF measurements (a day) is $1{\cdot}28 \times \widehat{\mathrm{SD}}$. Given that spirometric PEF in the ith child is stable at $\overset{\circ}{V}_i$ l/min, $\widehat{\mathrm{SD}}_i$ may be estimated directly using equation (7). Given that the same child has an estimated **m:s** gradient of $\hat{\beta}_i$, the absolute decline in **s** required to ensure that the expectation of **m** falls by at least 1·28 standard deviations from its stable value is $1{\cdot}28 \times \widehat{\mathrm{SD}}_i \div \hat{\beta}_i$. This therefore represents a realistic minimum for the magnitude of a true fall that has a reasonable chance of being detected within a day. Table V details the estimated magnitude of this fall for the 12 children in the study taking, as a convenient baseline, the observed mean spirometric PEF in the ith child as $\overset{\circ}{V}_i$. The table also details the percentage reduction from mean spirometric PEF that the estimated minimum fall in each child would represent. Figure 7 step [Q] details the MLn code which generates the values in Table V.

One may usefully contrast child 1 and child 3 (see Table V and Figure 9). Child 1 has a mean spirometric PEF of 437 l/min. At 437 l/min, $\widehat{\mathrm{SD}}$52·2 l/min, and in order for a true fall in meter-based PEF to be discriminated from random variation it must be at least $1{\cdot}28 \times 52{\cdot}2 = 66{\cdot}8$ l/min, but $\hat{\beta}_1 = 0{\cdot}1475$ and so in order to obtain a meter-based fall of 66·8 l/min it is necessary for spirometric PEF to fall by $66{\cdot}8 \div 0{\cdot}1475 \approx 453$ l/min which represents $453 \div 437 \times 100$ per cent ≈ 104 per cent of the baseline mean. Thus, in the case of child 1, it would seem that even a large fall in true PEF is unlikely to be detected by the portable meter. Child 3, on the other hand, has a mean spirometric PEF of 269 l/min. At 269 l/min, $\widehat{\mathrm{SD}} = 34{\cdot}4$ l/min and in order for a true fall in meter-based PEF to be discriminated from random variation it must be at least $1{\cdot}28 \times 34{\cdot}4 = 44{\cdot}0$ l/min, but $\hat{\beta}_3 = 0{\cdot}602$ and in order to obtain a meter-based fall of 44·0 l/min it is therefore necessary for true PEF to fall by $44{\cdot}0 \div 0{\cdot}602 \approx 73$ l/min, which represents ≈27 per cent of the baseline mean. Thus, in the case of child 3, there is a reasonable possibility that a clinically relevant fall in true PEF will be reflected in a recognizable fall in meter-based PEF. On the basis of this logic, it might reasonably be argued that portable PEF monitoring could be of potential value in children 2, 3, 5, 6, 7 and 11, but of little or no potential value in the other six children. These findings are reasonably robust to the choice of 1·28 standard deviations as the minimum detectable fall. If one chose to use 2 standard deviations as the cut-point, portable monitoring would still appear to be of potential value in children 2, 3, 6, 7 and 11, and of uncertain value in child 5.

2.5.3.4. Model criticism. Before drawing definitive conclusions it is necessary to subject the model to appropriate criticism.[26] The relationship between **m** and **s** was modelled as a straight line. This seems appropriate as the addition of square and cubic terms in **s** did not lead to an important improvement in model fit (see above). There was also no significant evidence of a requirement for additional terms in the random component of the model (see above).

Diagnostically standardized residuals (that is residuals divided by their *diagnostic* standard errors, see Goldstein, 1987, page 48[3]) were computed and plotted for all model terms (Figure 7 steps [R]-[T]). Level 1 standardized residuals for both **CONS**/**CONS** and **S**/**S** were approximately Normally distributed and exhibited no discernable relationship with fitted values generated using the fixed effects and level 2 residuals. The absolute magnitude of 5·9 per cent (expected 5 per cent) of both sets of standardized level 1 residuals was in excess of 1·96. The distribution of level 2 residuals was difficult to assess because of the small sample size (12 level 2 units only). One (8·3 per cent) **CONS**/**CONS** residual and no (0·0 per cent) **S**/**S** residual exceeded 1·96 in absolute magnitude. On the basis of a visual inspection of Figure 9, it appeared that


```
/* Read in data and put it into a SAS data set called 'pef'*/
data pef ;
infile 'pef.dat';
input child obs m cons s;
run;


/* Fit model (M4) using SAS Proc Mixed*/
proc mixed  data=pef  method=ml;
class=child;
model m=s / solution;
random intercept / solution g type=vc subject=child;
run;

/* Data=data set name.
   Method=ml specifies an analysis based on maximum likelihood (this is not the default).
   Class=child specifies that 'child' is a categorical variable.
   Model statement specifies 'm' as response variable and 's' as sole explanatory;
      a regression intercept is modelled by default (but can be removed as an option); 'solution'
      requests that parameter estimates are printed out.
   Random statement specifies that the regression intercept varies randomly between
      subjects and that subjects are defined by the 'child' variable; 'solution' requests that the
      child specific random effects (u_i and v_i) are printed out with their 'standard errors of
      prediction' which are equivalent to unadjusted comparative standard errors; g refers to the
      variance covariance matrix of the level 2 random parameters and vc specifies that it is
      diagonal (i.e. there is no covariance term, σ_uv=0)


   Fit model (M5) using SAS Proc Mixed
*/
proc mixed  data=pef  method=ml;
class=child;
model m=s / solution;
random intercept s/ solution g type=vc subject=child;
run;
/* Same as for model (M4) but effect of 's' specified as being random between subjects*/
```

/* denotes a comment in *SAS*.*/ Unbolded text represents active *SAS* code.

Figure 10. SAS Proc Mixed code for fitting models (4) and (5)

observations with a high regression leverage may have exerted unwarranted influence[3,26] upon the estimated gradients in children 1, 2, 5, 7, 9 and 11. The final model was therefore refitted on a data set from which these particular observations had been deleted. Basic conclusions were unaffected, the estimated overall gradient changed from 0·386 (see Table IV) to 0·412 [95 per cent CI 0·202 to 0·622] and the only child-specific change of note was that the gradient in child 11 became steeper (1·14). It was therefore concluded that the final model (Table IV) *did* provide a reasonable fit to the observed data.

2.5.3.5. Modelling in SAS. Figure 10 details the *SAS*[34] 'Proc Mixed' code required to fit models (4) and (5) and to generate child specific gradients from (5). The SAS estimated parameters for model (4) are very similar to those in Table III. The parameters of model (5) are also very similar

to those generated by MLn (data not shown): $\hat{\alpha}_0 = 61{\cdot}7$ (standard error = 25·14); $\hat{\beta}_0 = 0{\cdot}34$ (0·085); $\hat{\sigma}_v^2 = 7373$ (3113); $\hat{\sigma}_v^2 = 0{\cdot}063$ (0·036); $\hat{\sigma}_e^2 = 1867$ (107·4). The estimated level 2 random effects ($\hat{u}_i$ and $\hat{v}_i$) were also very close to their MLn counterparts. However, there was a consistent discrepancy in the estimates of the comparative variances; the comparative variances from MLn were up to 10 per cent larger than the square of the corresponding 'standard errors of prediction' from SAS. However, this is because the standard errors in SAS are not adjusted for the fact that the random parameters from which the residuals are estimated are themselves estimates.[3,30] When the residual type in MLn was changed to produce *unadjusted* comparative variances the results were very similar to those of SAS. As before, the gradient in the *i*th child was estimated as $\hat{\beta}_0 + \hat{v}_i$ generating values of 0·17, 0·47, 0·66, 0·13, 0·54, 0·53, 0·53, 0·10, 0·11, 0·09, 0·58 and 0·20, respectively.

2.5.3.6. Comparison with data resolution. By adding appropriate interaction terms to the analysis of covariance model (3), it was extended to allow for non-parallelism. The extended model $E(m_{ij}) = \alpha_i + \beta_i s_{ij}$; error = Normal) had a residual deviance of 1,128,383 on 605 degrees of freedom. The approximate likelihood ratio test statistic associated with the addition of the 11 additional fixed effects was $\chi^2_{11} = (1{,}224{,}320 - 1{,}128{,}383)/(1{,}128{,}383/605) = 51{\cdot}4$, $p < 0{\cdot}00001$. So, in keeping with the equivalent multi-level model-based test, this provides strong evidence that the **m:s** gradient *does* vary between children more than would be expected by chance. This suggests that an analysis of covariance which assumes parallelism may be misleading, and that an analysis which resolves each child's data to produce a single gradient and then uses that gradient as the raw data is perhaps the safest conventional approach.[2] The 12 estimated child-specific gradients are: 0·16; 0·56; 0·77; −0·01; 0·60; 0·73; 0·60; −0·09; 0·04; 0·02; 0·60; and 0·16, respectively. The empirical mean [95 per cent confidence interval] of these estimates is 0·35 [0·16 to 0·53] which is very similar to the overall gradient (0·386 [0·214 to 0·558]) estimated using the multi-level model (see Table IV). Furthermore, the individual child-specific gradients are qualitatively similar to their multi-level modelling equivalents (see Table V) although, as would be expected, they are generally more extreme relative to the overall mean. The only substantive change occurs in child 8; the child-specific gradient estimated by MLn is 0·34 while that estimated using the analysis of covariance model is −0·09. This is perhaps not surprising. Child 8 has the second smallest number of observations, the spirometric estimates of PEF are situated well above the population mean and there are two unusually low meter PEF values close to the middle of the range of spirometric PEF values. This all means that the estimated gradient in child 8 is likely to be particularly sensitive to different methods of estimating the corresponding intercept. This viewpoint is supported by fitting a GEE model with an exchangeable correlation structure, a single intercept and a separate gradient in each child (S-plus code detailed in Figure 11). This produces child-specific gradients of 0·16; 0·55; 0·78; −0·02; 0·62; 0·72; 0·60; 0·01; 0·04; 0·02; 0·63; and 0·17 with an empirical mean of 0·36 [0·18 to 0·54]. These results are qualitatively very similar to both the MLn model and the analysis of covariance, but the largest change from the analysis of covariance model occurs in child 8. In order to ensure that overall conclusions were not distorted by child 8, the MLn model was refitted with child 8 removed. The estimated mean gradient was 0·396 [0·209 to 0·583] and no child-specific gradients were markedly changed. This suggests that there was no serious distortion.

2.5.3.7. Clinical conclusions. The results of the quantitative analysis reported in this paper support the main conclusion of our earlier qualitative analysis.[1,7] They suggest that, in the clinical


```
# Data already read in and gee C program object code already linked (see Figure 3)

#  Create dummy variables allowing a separate gradient in each child
pef$s.dummy <- matrix(data=NA, nrow=629, ncol=12)
for(k in 1:12)
{
  pef$s.dummy[,k] <- (pef$child==k)*pef$s
}
# pef$s.dummy is first declared as a 629*12 matrix (i.e. a total of 629 observations in
# 12 children) with all elements set to missing (NA).
# The "for{}" loop then sets all elements in the k^th column to 0 except those relating to
# child k which take the value of the corresponding element of pef$s.

#  Fit model
gee.mod.slopes  <-  gee(x=pef$s.dummy, y=pef$m, id=pef$child,
                        varfun="gaussian", link="identity",
                        corstr="exchangeable", intercept=T, summarize=F)
# The only change from Figure 3 is that x (denoting the explanatory variables) is now
# declared to be the matrix pef$s.dummy rather than the single variable pef$s

#  To interpret output:
print(gee.mod.slopes$regest[2:13])
# Elements 2 to 13 of the vector gee.mod.slopes$regest now hold the slopes for individual
# children.

# Empirical mean of separate slopes
print(mean(gee.mod.slopes$regest[2:13]))

# Empirical standard error of empirical mean slope
print(sqrt(var(gee.mod.slopes$regest[2:13])/12))
```

 # denotes a comment in *Splus*. Unbolded text represents active *Splus* code.

Figure 11. S-plus code required to fit a GEE model with an exchangeable correlation structure allowing a separate slope in each child

setting, the response of portable PEF meters to changes in true PEF is relatively poor and that important falls in true PEF can be associated with declines in portable meter-based PEF that are markedly attenuated and might easily be obscured by random within-child variability. On the basis of the analysis described in this paper, we would now argue that although the overall response of portable PEF meters to changes in true PEF is poor, it is probable that this problem is more serious in some children than in others. This suggests that if it was possible to identify children in whom the response was good, regular portable PEF monitoring might well be useful in this subgroup.

3. ADDITIONAL ISSUES

3.1. Incomplete data

A number of approaches to the analysis of repeated measures data require balanced designs. One of the advantages of using GEEs[7] or multi-level modelling[3] is that provided the total number of

observations in each individual is random (in particular, provided it is independent of the value of the observed responses in that individual) an analysis based upon either approach will deal appropriately with the varying numbers of observations. There is therefore no need for designs to be balanced. For the same reason, studies which entail randomly missing response or explanatory data are also straightforward to analyse using these models.[3,7]

3.2. Interpreting regression coefficients

The motivation of the GEE model is the *marginal* distribution of the observed responses as a function of the covariates.[7,8,21] The predictive component of a linear regression model is called the *linear predictor* (in our GEE models it is $\alpha + \beta s_{ij}$) and in a marginal model it contains only conventional fixed effects;[5,7,8,20,29] the correlation or variance/covariance parameters are all estimated from the residuals. Because of this, regression coefficients are properly interpreted in a *population averaged* manner.[5,6] To illustrate, consider a covariate (B) with two levels designated *baseline* and *non-baseline*. The population averaged regression coefficient (β_{PA}) estimates the difference in expected response, having adjusted for all other covariates in the model, for the overall population of individuals as the non-baseline level of B compared to the overall population at the baseline level. This may be contrasted with β_{SS}, the subject specific regression coefficient, which estimates the change in the expected response for an *individual subject* if he/she was to move from the baseline level of B to the non-baseline level.[5,6] In the general case, for example in the GEE model which extends logistic regression, β_{PA} and β_{SS} can be quite different; in fact, if the repeated observations within an individual are positively correlated $\beta_{PA} \leqslant \beta_{SS}$.[5] However, when a response variable is *Normally* distributed and the link function is *identity* (as in simple linear regression) they are the same[29] and this interpretational issue is therefore unimportant in relation to the particular models we have considered in this paper.

Because multi-level modelling is generally used for Normally distributed responses the distinction between β_{PA} and β_{SS} is, again, usually irrelevant. However, readers should be aware that MLn supports some generalized models in which $\beta_{SS} \neq \beta_{PA}$.[30,36] When using such models one must make a deliberate choice between different approaches to model fitting which can modify the appropriate interpretation of regression coefficients and such modelling should not be undertaken unless one properly understands the relevant issues.[29,30]

3.3. Modelling a correlation structure which is of primary interest

Because GEE1 models do not provide a straightforward means of choosing between correlation structures, they are of limited use if it is the correlation structure which is of primary interest.[7] In contrast, provided the response is multivariate-Normal, multi-level modelling provides a powerful means of investigating a correlation or covariance structure. However, although MLn extends multi-level modelling to non-Normal responses (see above), some uncertainty exists as to the proper interpretation of the random effects estimated by MLn in such a setting[29] and if the response is not Normal and the correlation or covariance parameters *are* of primary interest a range of alternative approaches exist which may be preferable.[5,29] These include solutions based upon Markov chain Monte Carlo methods including Gibbs sampling.[37]

On the other hand, when it is the fixed regression parameters which are of primary interest and the correlation structure is merely a nuisance, GEE1 models can be invaluable, particularly if the response is non-Normally distributed. GEE1 models are commonly used to analyse rates (or counts) which are *Poisson* distributed and proportions (including binary [yes/no] responses)

which are *binomially* distributed.[5–8] Some standard GEE1 programs also support *gamma* and *inverse Gaussian* models.

4. CONCLUDING REMARKS

Studies with designs based upon repeated measures are an essential feature of much of the research in epidemiology and clinical science. Such designs can be very powerful, both statistically and scientifically, because they enable one to study changes *within* individuals over time or under a variety of different conditions. However, their power arises directly from the fact that repeated measurements tend to be correlated with one another, and this correlation *must* be addressed at the time of analysis. Failure to take proper account of a correlation structure can lead to conclusions that are *qualitatively incorrect*. It is therefore important for researchers in epidemiology and clinical science to be aware of the rich variety of methods that now exist to deal with the correlation structures that can arise from designs such as these. These approaches include methods based upon generalized estimating equations and models incorporating random effects, including multi-level modelling. We hope that this tutorial has demonstrated that these methods have a logical basis which can be understood by non-statisticians and that in simple cases they produce results which are reassuringly similar to more conventional approaches. Nevertheless, we also hope that we have primed the reader to recognize that in more complex settings there are a number of interpretational pitfalls and it is essential that the user understands every model that he or she creates and seeks appropriate advice if clarification is required. Potential users should be prepared to spend a considerable amount of time coming to grips with the required computer packages and in ensuring that they properly understand the models which they ultimately fit.

APPENDIX: STANDARD GEE CORRELATION STRUCTURES

Independence

Given that a data set consists of repeated measurements within individuals, the simplest possible correlation structure is known as *independence*. This is equivalent to making the assumption that each observation in an individual is completely uncorrelated with every other observation in that individual. If ρ_{jk} is the correlation between observations j and k, $\rho_{jj} = 1$ and $\rho_{jk[j \neq k]} = 0$.

Exchangeable or Compound Symmetry

Every observation within an individual is equally correlated with every other observation in that individual. Formally, $\rho_{jj} = 1$ and $\rho_{jk[j \neq k]} = \rho$ where ρ is the *intraclass correlation coefficient*.

Autoregressive

An autoregressive correlation structure indicates that two observations taken close in time (or space) within an individual tend to be more closely correlated than two observations taken far apart in the same individual. Formally, $\rho_{jj} = 1$ and $\rho_{jk[j \neq k]}$ increases in value as the absolute difference between j and k falls. As a specific example, a first-order autoregressive (AR-1) correlation structure specifies that $\rho_{jk} = \rho^{|j-k|}$ where ρ is the correlation when $|j - k| = 1$.

Unstructured

This is used in balanced data sets. No assumption is made about the relative magnitude of the correlation between any two pairs of observations. Formally, $\rho_{jj} = 1$ and $\rho_{jk[j \neq k]}$ is free to take any value between -1 and $+1$.

User fixed

All correlation coefficients are fixed by the user rather than being estimated from the data. Formally, $\rho_{jj} = 1$ and $\rho_{jk[j \neq k]}$ can take any value between -1 and $+1$, but this value is fixed prior to the analysis rather than being estimated from the data.

ACKNOWLEDGEMENTS

We must first thank Patricia Cahill RN who supervized every peak flow measurement; without her it would have been impossible to conduct the research upon which our practical example was based. We would also like to thank all of the boys who participated in the study for their time, effort and patience. We thank the statisticians, epidemiologists and clinical scientists who commented on earlier versions of the manuscript. We gratefully acknowledge the teaching and advice provided by the Multilevel Models Project Team at the Institute of Education, London. This research was supported in part by grants from the National Health and Medical Research Council of Australia.

REFERENCES

1. Armitage, P. and Berry, G. *Statistical Methods in Medical Research*, 2nd edn, Blackwell Scientific Publications, Oxford, 1987, pp. 70, 82, 86, 90, 104–106, 109, 143–150, 156, 273–295, 302, 307, 314–316, 318, 321, 347–357, 506.
2. Feldman, H. A. 'Families of lines: random effects in linear regression analysis', *Journal of Applied Physiology*, **64**, 1721–1732 (1988).
3. Goldstein, H. *Multilevel Models in Educational and Social Research*, Charles Griffin & Company Ltd, London, 1987, pp. 1, 10–31, 32–33, 48, 51–60, 83–84.
4. Laird, N. 'Longitudinal studies with continuous responses', *Statistical Methods in Medical Research*, **1**, 225–247 (1992).
5. Neuhaus, J. 'Statistical methods for longitudinal and clustered designs with binary responses', *Statistical Methods in Medical Research*, **1**, 249–273 (1992).
6. Zeger, S. L. and Liang, K-Y. 'An overview of methods for the analysis of longitudinal data', *Statistics in Medicine*, **11**, 1825–1839 (1992).
7. Liang, K-Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
8. Zeger, S. L. and Liang, K-Y. 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, **42**, 121–130 (1986).
9. Steimer, J-L., Mallet, A. and Mentre, F. 'Estimating interindividual pharmacokinetic variability', *in* Rowland, M., Sheiner, L. B. and Steiner, J-L. (eds), *Variability in Drug Therapy*, Raven, New York, 1985, pp. 95–101.
10. Goldstein, H. 'Multilevel mixed linear modelling analysis using iterative generalized least squares', *Biometrika*, **73**, 43–56 (1986).
11. Gardner, R. M., Crapo, R. O., Jackson, B.R. and Jensen, R. L. 'Evaluation of accuracy and precision of peak flowmeters at 1,400m', *Chest*, **101**, 948–952 (1992).
12. Miller, M. R., Dickinson, S. A. and Hitchings, D. J. 'The accuracy of portable peak flow meters', *Thorax*, **47**, 904–909 (1992).
13. Miller, M. R. and Pedersen, O. F. 'The peak flow working group of the European Respiratory Society. The characteristics and calibration of devices for recording peak expiratory flow', European Respiratory Society, Birmingham, 1993.

14. Clark, N. M., Evans, D. and Mellins, R. B. 'Patient use of peak flow monitoring', *American Review of Respiratory Diseases*, **145**, 722–725 (1992).
15. Harm, D. L., Kotses, H. and Creer, T. L. 'Portable peak-flow meters: intrasubject comparisons', *Journal of Asthma*, **21**, 9–13 (1984).
16. Lloyd, B. W. and Ali, M. H. 'How useful do parents find home peak flow monitoring for children with asthma? *British Medical Journal*, **305**, 1128–1129 (1992).
17. Sly, P., Cahill, P., Willet, K. and Burton, P. 'The accuracy of mini peak flow meters following changes in lung function in asthmatic children', *British Medical Journal*, **308**, 572–574 (1994).
18. Altman, D. G. and Bland, J. M. 'Measurement in medicine: the analysis of method comparison studies', *Statistician*, **32**, 307–317 (1983).
19. Bland, J. M. and Altman, D. G. 'Statistical methods for assessing agreement between two methods of clinical measurement', *Lancet*, **i**, 307–310 (1986).
20. Zeger, S. L., Liang K-Y. and Albert, P. S. 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics*, **44**, 1049–1060 (1988).
21. Liang, K.-Y., Zeger, S. L. and Qaqish, B. 'Multivariate regression analysis for categorical data', *Journal of the Royal Statistical Society*, *Series B*, **54**, 3–40 (1992).
22. Hanfelt, J. 'GEE4 documentation', Technical report, John Hopkins University, 1993.
23. Chambers, J. M. and Hastie, T. J. *Statistical Models in S*, Wadsworth and Brooks, Pacific Grove, 1992.
24. *Splus. User's Manual*, Statistical Sciences Inc, Seattle, 1991.
25. Carey, V. J. *C Implementation of gee for S (92/8/7)*, Harvard University, Boston, 1992.
26. McCullagh, P. and Nelder, J. A. *Generalized Linear models*, 2nd edn, Chapman and Hall, Oxford, 1989, pp. 21–47, 404–409, 471.
27. Aitkin, M., Anderson, D., Francis, B. and Hinde, J. *Statistical Modelling in GLIM*, Clarendon Press, Oxford, 1989, pp. 76, 80, 146.
28. Huber, P. J. 'The behaviour of maximum likelihood estimates under non-standard conditions', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 221–233 (1967).
29. Breslow, N. E. and Clayton, D. G. 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, **88**, 9–25 (1993).
30. Rasbash, J. and Woodhouse, G. *MLn Command Reference. Version 1.0.*, Institute of Education of the University of London, London, 1995, pp. 43–45, 80.
31. Prosser, R., Rasbash, J. and Goldstein, H. *ML3, Software for Three-level Analysis*, Institute of Education of the University of London, London, 1991, pp. 12, 18–20.
32. Healy, M. J. R. *Nanostat Users' Guide*, London School of Hygiene and Tropical Medicine, London, 1987.
33. *Minitab Reference Manual*, Minitab Inc, State College of Pennsylvania, 1989.
34. SAS Institute Inc. *UNIX release 6.09*, SAS Institute Inc, Cary, 1993.
35. Groemping, U. *GEE: A SAS Macro for Longitudinal Data Analysis*, Universitaet Dortmund, Dortmund, 1994.
36. Goldstein, H. 'Nonlinear multilevel models, with an application to discrete response data', *Biometrika*, **78**, 45–51 (1991).
37. Zeger, S. L. and Karim, M. R. 'Generalized linear models with random effects; a Gibbs sampling approach', *Journal of the American Statistical Association*, **86**, 79–86 (1991).

