WHY NOT ROUTINELY USE BEST LINEAR UNBIASED PREDICTORS (BLUPs) AS ESTIMATES OF CHOLESTEROL, PERCENT FAT FROM KCAL AND PHYSICAL ACTIVITY?

EDWARD J. STANEK III*, ARNOLD WELL2 AND IRA OCKENE3

1 404 Arnold House, Department of Biostatistics and Epidemiology, University of Massachusetts at Amherst, MA 01002, U.S.A.
2 Department of Psychology, University of Massachusetts at Amherst, U.S.A.
3 Department of Medicine, University of Massachusetts at Worcester, U.S.A.

SUMMARY

Measures of biologic and behavioural variables on a patient often estimate longer term latent values, with the two connected by a simple response error model. For example, a subject's measured total cholesterol is an estimate (equal to the best linear unbiased estimate (BLUE)) of a subject's latent total cholesterol. With known (or estimated) variances, an alternative estimate is the best linear unbiased predictor (BLUP). We illustrate and discuss when the BLUE or BLUP will be a better estimate of a subject's latent value given a single measure on a subject, concluding that the BLUP estimator should be routinely used for total cholesterol and percent kcal from fat, with a modified BLUP estimator used for large observed values of leisure time activity. Data from a large longitudinal study of seasonal variation in serum cholesterol forms the backdrop for the illustrations. Simulations which mimic the empirical and response error distributions are used to guide choice of an estimator. We use the simulations to describe criteria for estimator choice, to identify parameter ranges where BLUE or BLUP estimates are superior, and discuss key ideas that underlie the results. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Many clinics/medical centres have large amounts of data on biologic parameters for patients. Although an individual patient's data are used at a patient's visit, rarely are other assembled data used to improve the patient's estimate. Instead, clinicians use patient's data to estimate individual patient parameters. An important part of clinical practice is establishing protocols for measurement of these parameters so that they are sufficiently accurate to be biologically meaningful. For example, fasting periods are commonly required before cholesterol measures are made, and replicate measures are often made to reduce error. Although attention is given to ensure the accuracy of the patient-specific measures, gains in accuracy are achieved primarily by setting protocols, and replication of measures on the patient.

Although other assembled data are rarely used to improve the accuracy of a patient's estimate, it could be argued that improved accuracy of an individual patient's estimate is possible by using
previously collected data from other clinic patients. However, some may claim that such data should not be used, since their use will introduce bias, and ultimately result in less accurate estimates. The purpose of this paper is to discuss how estimation of a patient’s parameter can be improved by using data from other patients. In particular, we illustrate that BLUP estimates are better than BLUE estimates for two of three common patient parameters. This paper is primarily expository, although the discussion draws in key theoretical references that underpin the results.

We develop these issues in the context of a large observational longitudinal study of seasonal variation in cholesterol levels which we refer to as the Seasons study (Seasonal Variability of Blood Lipids, NHLBI, number R01-HL52745). Data in this study were collected on a volunteer sample of enrollees between the ages of 20 and 70 in the Fallon Health plan (a Massachusetts’s health maintenance organization) with the goal of identifying and quantifying factors (such as food intake, exercise, light exposure and weight) that relate to seasonal changes in cholesterol. Measures of each participant’s cholesterol, food intake and exercise were made at five consecutive three-month intervals. Our focus is not on seasonal effects, but rather on the properties of alternative estimators of patient parameters (defined as average daily values over a one-year time period). Results from the Seasons study data are used to set the context for comparison of estimators, using as examples total cholesterol, per cent kcal from fat, and total resting metabolic rate (per hour) equivalent from leisure time activities (MET).

We first define models for patient-specific parameters (called latent values) in the context of the Seasons study and use the basic definitions to describe a fixed-effect and mixed model for the data. In the fixed-effect model, estimates of patient parameters are typically BLUE estimates from ordinary (or weighted) least squares, while in the mixed model, estimates of realized random effects are typically BLUP estimates. We focus on a comparison of these two estimators, distinguishing between the patient’s latent value and the expected value of the observed response (the linear predictor). Using this distinction in target parameters, we identify the appropriate target parameter for a particular patient, and compare via simulation BLUE and BLUP estimators over a range of patient-specific parameter values.

We begin by studying properties of these estimators using data for total cholesterol from the Seasons study. First, we describe the latent distribution of total cholesterol using average subject measures, and the response error distribution for the subjects. Then, using empirical representations of these distributions, we simulate a single measure of cholesterol on a sampled subject, and use the simulation to compare the BLUE and BLUP estimates via their mean squared error (about the subject’s true latent value). We summarize this comparison based on average results conditional on a subject’s true latent value, and conditional on results given an observed value.

Finally, the results are presented in settings where variance components are estimated from the data. Similar results are presented for two other response variables, per cent kcal from fat (based on 24-hour recall telephone interviews), and leisure time activity METs (based on 24-hour activity recall). The discussion integrates these results, and describes limitations and implications for their use.

2. SIMPLE MODELS FOR PATIENT PARAMETERS

We first define latent value parameters for patients in a population. These parameters are non-stochastic, with a single parameter for each patient in the population listing. We use these parameters in the definition of models that account for sampling of patients, and sampling of
measures on a patient. Finally, we define competing parameter estimates and evaluate their properties in these models.

We represent the parameter for patient \( s \) on day \( d \) by \( \mu_{sd} \). For discrete variables such as the number of meals eaten in a day, \( \mu_{sd} \) may be observed directly without ‘error’. For other variables, such as total cholesterol, a measure of total cholesterol is likely to differ from \( \mu_{sd} \) by an amount \( \epsilon_{sd}^* \), where \( r \) indexes the difference which may be due to temporal variation in total cholesterol over the day, variation due to sampling of blood, and laboratory measurement error. The quantity \( \epsilon_{sd}^* \) is the realized value of the random variable \( E_{sd}^* \). We define \( \mu_{sd} \) as the expected value of total cholesterol on day \( d \) for patient \( s \), resulting in the expected value of \( E_{sd}^* \) over potentially repeated measures on the patient-day being zero, and define the variance of \( E_{sd}^* \) by \( \sigma_{sd}^2 \). Thus, our definition implies that the measured value of a variable is an unbiased estimate of \( \mu_{sd} \).

We define the parameter \( \mu_s \) for patient \( s \) to be the simple average over a specified time period (365 days) of the daily patient parameters. This parameter is the latent value for the subject. The choice of a 365 day time period is arbitrary; different time periods may be defined that have different clinical significance. For simplicity, we limit our discussion to settings where measures on a patient result from a random sample of days, with a single measure of response on each day. For the \( k \)th measure, we represent response on patient \( s \) by \( \mu_s + E_{sk} \), noting that the random term \( E_{sk} \) (which we call response error) incorporates both day to day variation and other sources of error that contributed to \( \sigma_{sd}^2 \). Our assumptions and definitions imply that the expected value of \( E_{sk} \) is zero. We represent the variance of \( E_{sk} \) by \( \sigma_{se}^2 \).

We use these definitions of parameters for patients to define parameters for a population of \( s = 1, \ldots, M \) patients, where the average patient parameter is given by \( \mu \), and the variance of patient parameters is given by \( \sigma_B^2 \). We define \( \sigma_{sp}^2 \) using a divisor of \( (M-1) \) rather than \( M \) to simplify formula, but assume that \( M \) is large enough that this distinction is not important. We represent the parameter for patient \( s \) in terms of the overall population mean and a patient effect, \( \beta_s \), where \( \mu_s = \mu + \beta_s \). With these assumptions and definitions, a model for response on the \( k \)th selected day for patient \( s \) is given as

\[
Y_{sk} = \mu + \beta_s + E_{sk}
\]

where the expected value over response error, \( E_R(Y_{sk}) = \mu + \beta_s = \mu_s \), is the latent value for subject \( s \), and the variance over response error is given by \( V_R(Y_{sk}) = \sigma_{se}^2 \). This model is commonly described as a fixed effect model for patients. Often, response error will be assumed to be homogeneous over patients. Best linear unbiased estimates of latent values for patients are ordinary least squares estimates, and correspond to the simple average of the observed responses for each patient

\[
p_1 = \bar{y}_s.
\]

If only one measure is available on a patient, this estimate simplifies to the observed response.

We distinguish model (1) representing response for patient \( s \) in the population from a model for response from the \( i \)th selected patient from the population given by

\[
Y_{ik} = \mu + B_i + E_{ik}.
\]

We draw attention to the difference in interpretation by replacing the subscript \( s \) (designating the \( i \)th selection of a patient) and by reducing the deviation for patient \( s \) by the deviation for the \( i \)th selected patient, \( B_i \). While \( \beta_s \) is a fixed effect in model (1), \( B_i \) is
a random effect in model (2). \(B_i\) is random, not because the effect for the particular selected patient (the realized patient) is unknown, but rather because we do not know which patient will be selected. More formally, defining a set of indicator random variables \(S_{is}\) having value of 1 if patient \(s\) is selected on the \(i\)th selection, and 0 otherwise for \(s = 1, \ldots, M\), we can express the random effect as a linear combination of the sampling indicator variables times the fixed effects, such that

\[
B_i = \sum_{s=1}^{M} S_{is} \beta_s
\]

When all \(S_{is}\) (for \(s = 1, \ldots, M\)) are realized, the resulting parameter is a realized random effect. Similar expressions can be constructed for \(Y_{ik}\) and \(E_{ik}\) in terms of \(S_{is}\) for \(s = 1, \ldots, M\) and \(Y_{sk}\) and \(E_{sk}\), respectively.\(^1,2\)

Model (2) is usually described as a random effect model, but can also be considered to be the simplest example of a mixed model (with only a single fixed parameter). Historically, estimation for the random effects model has been focused on the variance of the random effect. More recently in the context of mixed models,\(^3-7\) methods have been established and popularized for estimating realized random effects. The estimators are commonly called best linear unbiased predictors (BLUPs). With known variance components, the BLUP estimates for the \(i\)th selected subject is given by

\[
p_2 = \bar{y} + k_i (\bar{y}_i - \bar{y})
\]

where

\[
k_i = \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_\epsilon/n_i}
\]

with \(n_i\) representing the number of measures on the \(i\)th selected subject, and \(\sigma^2_\epsilon\) representing the expected value of response error over patients in the population. This estimate differs from the BLUE estimate for patient \(s\).

The simplest comparison of the BLUP and BLUE estimate occurs when there is only one measure on a subject, and the overall sample size is large enough so that we can approximate \(\bar{y} \approx \mu\).

With this approximation, and assuming the \(i\)th selected subject is subject \(s\), \(p_1 = y_{sk}\), while \(p_2 = \mu + k_i (\bar{y}_{ik} - \mu)\). Since whenever \(\sigma^2_\epsilon > 0, k_i < 1\), the BLUP estimator will be closer to the mean than the BLUE estimator. The amount that the BLUP estimator will be closer to population mean depends on \(k_i\), which when \(n_i = 1\) (which we subsequently assume) corresponds to the intraclass correlation, or reliability.\(^8\) For variables that have large response variance relative to subject variance (low values of \(k\)), the BLUP estimates will differ greatly from the observed response, and be much closer to the population mean.

Terminology for BLUP and BLUE estimates is somewhat confusing since the acronyms for both estimators include ‘U’ representing unbiased. The potential confusion results from different definitions of bias. For the BLUE estimator, the traditional definition of unbiasedness is used such that \(E_R(p_1) = \mu_s\). Hence the BLUE estimate of response for patient \(s\) is unbiased. For BLUP estimates, unbiasedness refers to the average bias over all possible selection of patients. Letting \(S\) denote expectation over patient selection, unbiasedness for BLUP estimates implies that \(E_S (p_2) = E_S [E_{RS} (\mu + k (\bar{y}_{ik} - \mu)))] = E_S [\mu + k (\mu - \mu)] = \mu\). Given that the \(i\)th subject selected
is subject \( s, E_R[p_2] = \mu_s + (k - 1)(\mu_s - \mu) \). The second term in this expression represents the bias of the BLUP estimate of a realized subject’s latent value.

There is one additional opportunity for confusion when discussing the bias of BLUP estimates. The potential confusion stems from different possible definitions of the target parameter that is the goal of the estimation. The previous discussion assumes that the target parameter is the latent value for the subject given by \( \mu_s \). Suppose that we assume that the target parameter is given by \( E_S[\mu + B_i|Y_{ik}] \), which is called the linear predictor.\(^9\) Given \( Y_{ik} \), this expected value will equal a linear combination of \( B_i \) where the weighting of the \( B_i \) will be proportional to the probability that \( E_i = (Y_{ik} - \mu) - B_i \). When \( B_i \) is normally distributed, and independent of \( E_i \) (also normally distributed) and \( n_i = 1 \) for all \( i \), then \( E_S[\mu + B_i|Y_{ik}] = \mu + k(Y_{ik} - \mu) \). Replacing \( Y_{ik} \) by a realized value, \( y_{ik} \), in this expression produces a value identical to the BLUP estimate. Under these assumptions, if the target parameter is \( E_S[\mu + B_i|Y_{ik}] \), then the expected value of the BLUP estimator is equal to the linear predictor. In this sense, the BLUP estimator is unbiased.

Since the BLUP estimates of realized random effects are biased estimates of the latent value, their comparison with BLUE estimates is based on the mean square error formed by taking the expected value of the squared deviation of the estimate from the latent value. The estimator with the smallest MSE we consider to be best. Since the bias increases proportionally to the distance between the latent value and the population mean, no single estimator is uniformly best over the range of latent values. Stanek shows that for large \( M \), the BLUP estimates will have smaller MSE whenever \( \beta^2 < \sigma^2 + 2\sigma^2_b \), regardless of the distribution of latent values or response error.\(^1\) This conclusion is of limited value since the choice of an estimator (in terms of minimum MSE) is based on knowledge of the parameter one is trying to estimate. A more practical question, and one that we address empirically here, is which estimator is better given the observed response.

### 3. CONTEXTS FOR COMPARING THE MSE OF ESTIMATORS

The MSE of an estimator is defined as the expected squared deviation between the estimate and the realized latent value, where expectation is calculated over repeated samples. For an observed response on a patient, there are two ways that these repeated samples can be constructed. First, we can imagine a sequence (A) of single measurements on the patient, with the sequence of measures corresponding to independent samples. Alternatively, we can imagine the collection of patients who could have had the observed response as constituting a population of patients (with patients represented in the population proportional to the chance that they could have had the observed response), and that the observed patient has been randomly selected from this population. We refer to a sequence of independent selections of patients from this population as sequence (B), where the response observed on the selected patient is the measured response (which is identical for all patients). The first sequence is associated squarely with a frequentist based approach to inference; the patient’s latent value (parameter) does not vary in the sequence, but the imagined set of observed responses do vary. The expected value is formed by averaging the sequence of observed values (a process identical to the operational strategy of obtaining multiple measures on the patient). The second sequence is associated with a likelihood based, or Bayesian approach to inference, with the observed response not varying in the sequence, but the imagined set of latent values (parameters) varying. The expected value of this sequence is formed via a weighted average of the sequence of latent values, with weight equal to the probability of the latent values, with weight equal to the probability of the latent value. (At the risk of confusing the
issue, this second sequence of samples underlies the notion of ‘regression to the mean’, an idea embraced by both frequentist and Bayesians alike.)

The manner in which the sequence of repeated samples is constructed is important when considering the bias of the BLUE and BLUP estimators. When evaluated over sequence A, the BLUE estimate is an unbiased estimate of the patient’s latent value. Over this same sequence, the BLUP estimate is biased, underestimating the patient’s latent value when the observed measure is above $\mu$, and overestimating the patient’s latent value when the observed measure is below $\mu$. When considered over sequence B, the BLUE estimate is identically equal to the observed value, and is a biased estimate of the (weighted) average latent value for patients in the sequence, but not generally equal to the specific patient’s latent value. Under sequence B, the BLUP is an unbiased estimate of the (weighted) average of the latent values for patients in the sequence, but is not unbiased for a specific patient’s latent value. When considering bias as a criteria for estimation, only the BLUE estimate under sequence A produces an unbiased estimate of the patient’s latent value. This result, plus the traditional emphasis on unbiasedness, has focused attention on sequence A and the BLUE estimator in practice.

Requiring unbiasedness in estimation limits consideration of alternative estimators. Relaxing this criterion allows consideration of a broader set of estimators, some of which may come closer in terms of average squared distance from the latent value (that is, MSE) than the BLUE estimator. Thus, we drop our focus on the absence of bias, and consider the MSE.

4. EMPIRICAL COMPARISONS OF BLUE AND BLUP ESTIMATORS

We study the MSE of BLUE and BLUP estimates of realized random effects using simulations based on an empirical distribution of latent values and response error constructed from the Seasons study. We repeat the simulation using several response variables (total cholesterol, per cent kcal intake from fat, and leisure time activity (in METs)) selected based on their different parameter and response distributions. Our goal is to compare properties of BLUP and BLUE estimates of patient parameters in a realistic context over variables with different response distributions.

For each variable, the simulation followed a similar process. First, an empirical distribution of average subject response (which served as the distribution of the parameters $\mu_k$) was constructed using subjects that met minimum data requirements ($\geq 4$ cholesterol measures, or $\geq 10$ per cent kcal from fat or MET measures) from the Seasons study. Subtracting the subject’s average from the subject’s response to obtain $E_{sk}$, and pooling results over subjects, an empirical response error distribution was also tabulated. Since the number of measures per subject differed, we characterized the empirical distributions by an estimate of the overall population mean and mixed model REML variance component estimates for the subject and response error variance components (using SAS Proc Mixed®).

We used the subject means from the Seasons study to construct a distribution of latent values by determining the proportion of subjects falling in equal width class intervals, and used the same interval widths to define the response error distribution. Class intervals were used so that subjects could be grouped for comparison of MSE between BLUP and BLUE estimates of latent value. For total cholesterol, the latent distribution was constructed from $n = 505$ subjects classified in 29 intervals with a width of 10 mg/dl; for per cent kcal from fat, the latent distribution was constructed from $n = 493$ subjects classified in 38 intervals of width per cent; for MET total in leisure time activity, the latent distribution was constructed from $n = 470$ subjects classified in 70 intervals of width 0.1 MET/d.
We simulated subjects’ responses by selecting a latent value from a class interval using the interval’s probability, and adding to it a response error selected in a similar manner from the response error distribution. We made the simplifying assumption that response error was independent of the latent value. This assumption was supported by the weak relationship between the subject specific mean and variance (or standard deviation) for variables. Within a selected class interval, we used a uniform random number generator to choose a particular value. For each subject’s latent value selected, we generated two responses, with the purpose of being able to estimate the mean and shrinkage constant from the empirical data. A total of 25,000 latent values were selected, generating a total of 50,000 responses for each simulation.

For each simulated response, the BLUP estimate was calculated, and the MSE (relative to the latent value) for the observed response and the BLUP estimate evaluated. These MSEs were averaged over class intervals defined by subject latent values, and over class intervals defined by subject’s realized responses. The difference in the MSEs was plotted to form a summary comparison of the results.

We evaluated the impact of using estimates of variance components to form BLUPs in two ways. First, we defined a study size and ‘study’ by using consecutive groups of $m$ subjects (where $m = 5, 25$ and $100$). Since the simulation generated two measures for each subject, we used the simulated data to estimate the overall mean, and variance components for the subject and response variance (replacing negative estimates of subject variance by 0). Empirical BLUP (EBLUP) estimates were constructed using these estimated parameters in place of the population parameters. In the first evaluation, we used the estimated parameters for the same set of subjects that generated the parameters. For the second evaluation, we used the parameters estimated from a particular set of ‘study’ subjects to form EBLUP estimates for the next consecutive set of selected subjects (with estimates from the last study set being used for subjects in the first study). We present results comparing EBLUP estimates based on consecutive groups of $m = 5$ subjects (and $m = 25$ for total cholesterol), noting that results based on larger group sizes fell intermediate between these results and results based on true simulation parameters.

5. RESULTS

We compare the BLUP and BLUE estimators for each response variable via a set of figures. The first two figures in each set are histograms of the simulated observed response distribution, and the simulated distribution of latent values. The third figure summarizes the average difference (BLUE − BLUP) in MSE between the estimators based on groupings corresponding to subject’s latent values, and grouping corresponding to subject’s observed response. The MSE comparisons are smoothed to facilitate comparison. Results are included based on using empirical BLUPs for study sets of $m = 5$ subjects (Figures 3(a), 6 and 9) for each response variable, and $m = 25$ subjects (Figure 3(b)) for total cholesterol. The figures compare the MSE when the estimated mean and shrinkage constant is applied to the same subjects, and the estimates from the previous study set are used to estimate variances in forming BLUP estimates for the next study set of subjects.

The mean cholesterol among the 505 Seasons study subjects based on 2424 measures was 218 mg/dl, with estimates of $\sigma_p^2$ and $\sigma_e^2$ given by 1530 and 308, respectively, resulting in $k = 0.84$. For the simulation, the mean cholesterol per subject was taken to be the subject’s latent values. The simulated cholesterol latent distribution was slightly skewed to the right (Figures 1 and 2) with a discontinuity in the right tail of the latent value distribution. For a subject with an...
observed serum cholesterol measure of 245 mg/dl, the BLUP estimate is 240 mg/dl based on the population parameters. Since 240 mg/dl is often used as a screening level for elevated serum cholesterol, use of a BLUP estimate instead of the observed response (BLUE estimate) would change the interpretation of results for approximately 4 per cent of the subjects. Figure 3(a) plots the average difference in MSE (observed minus BLUP) versus serum cholesterol levels, based on 10 mg/dl serum cholesterol groups. The concave solid curve is formed by grouping responses according to subject’s latent values with BLUP estimates evaluated using known variance parameters. The remaining curves represent the average difference in MSE using total cholesterol groupings constructed from 10 mg/dl observed cholesterol groups. The two convex curves in the upper half of Figure 3(a) are formed assuming variance components are known, or by using estimated variance components (based on \( m = 5 \)) to construct BLUP estimates for the observations used to estimate them. The widely fluctuating curve in the lower half of Figure 3(a) is based on using the previous study set estimated variance components for the current study set of subjects (with \( m = 5 \)).
All of the curves are constructed from a simulation of two measures on each of 25,000 subjects. Points plotted for cholesterol value intervals between 120 and 310 mg/dl are based on the average MSE for a minimum of 100 simulated observations. Points falling above the reference line of zero have smaller MSE for BLUP estimates, as compared with BLUE estimates. Although the tabulation of MSE is based on observed or latent cholesterol values, the MSE for each estimator is always calculated relative to the subject’s latent value. The concave curve indicates that when grouping subjects by their true latent cholesterol levels, the BLUP estimates are not uniformly best. Using the results of Stanek,¹ BLUP estimates will have smaller MSE when latent cholesterol values are between 160 and 276 mg/dl, a result that occurs for 86 per cent of the subjects. Since BLUP estimates are biased given a subject’s observed value with the bias increasing with increasing distance of the latent value from the population mean, for subjects with unusually large (and small) latent cholesterol levels, BLUP estimates have larger MSE than the observed cholesterol measure.

The MSE comparison when tabulated based on subject groupings formed by the observed serum cholesterol values form a different picture. The convex curve in Figure 3(a) indicates that the BLUP estimates have smaller MSE for nearly all responses (99.57 per cent) as compared to the observed response (BLUE estimate), when the subject’s true variance parameters are known. The few observations where the average MSE difference favours the observed response occurs for observations near the discontinuity in latent values. The results indicate that using a BLUP estimate as opposed to the observed cholesterol level will come closer to the true subject’s cholesterol level nearly all the time. Although for most observations the magnitude of the reduction is small, the consistent results motivate BLUP estimates as the default.

A similar results occurs when empirical BLUP estimates are calculated using the mean and variance estimates based on study sets of \( m = 5 \) subjects, and the estimates are used for subjects in the study set used to calculate the variance. Since variance parameters are calculated based on only five subjects, the estimates have high variance. This large variance is translated into slightly larger MSE for empirical BLUP estimates in the central range of total cholesterol (from 170 to 230 mg/dl), but a similar reduction in MSE for BLUP estimates for other observed cholesterol levels. When the mean and variance estimates from the previous study set of subjects are used to calculate the EBLUPs for the next study set of five subjects, the EBLUP estimates have much

---

Figure 3. Difference in MSE (observed TC – BLUP) by observed latent total cholesterol (mg/dl) for: (a) study size 5; (b) study size 25

---

²⁹⁵¹

BEST LINEAR UNBIASED PREDICTORS AS ESTIMATES OF BIOLOGIC PARAMETERS	 ²⁹⁵¹

Copyright © 1999 John Wiley & Sons, Ltd.

larger MSE than using the observed cholesterol as an estimate. This comparison of MSEs indicates that for cholesterol, a practical lower limit will be reached where the lack of precision in variance and the mean estimates will overwhelm any reduction in MSE theoretically expected to have occurred.

We evaluated the sensitivity of MSE comparison to the estimated parameters used in forming the BLUP estimates by repeating the simulations using different study set sizes. Figure 3(b) contains results similar to Figure 3(a), but based on \( m = 25 \) subjects in a study set. Notice that when the mean and variance parameters are estimated using \( m = 25 \) subjects, the empirical BLUP estimates have smaller MSE for nearly all observed data, even when the estimated parameters are applied to construct BLUP estimates in the next study set of subjects.

Similar results occurred when estimating the per cent of kcals that was due to fat based on 24-hour food recalls. A total of 493 subjects completed 6859 24-hour dietary records over a one year time period, with a minimum of 10 dietary records per person. Dietary recalls were obtained by random telephone interviews on two week days and one weekend day during four consecutive three-month periods. Food quantities and items were converted to kcal and nutrient amounts with the per cent kcal from fat estimated for each subject-day. The mean per cent kcal from fat was 31.3 per cent, with subject and response variance of 34.0 and 60.0, resulting in \( k = 0.36 \).

Figures 4 and 5 present histograms of the simulated empirical latent distribution of per cent kcal from fat, and response error, while Figure 6 summarizes the comparison of MSE of BLUE and BLUP estimates based on a single 24-hour dietary recall for a subject. The concave curve indicates that when the true per cent kcal from fat intake is between 18 and 41 per cent (which occurred for 86.2 per cent of the subjects), the BLUP estimate had smaller MSE. For subjects with true latent values outside of this range, the observed per cent kcal from fat intake for a single 24-hour recall provided a more accurate estimate of the subject’s true per cent kcal from fat intake. A more interesting question concerns comparison of estimates of a subject’s latent per cent kcal from fat intake given the subject’s observed per cent kcal from fat intake.

Three convex curves are included in Figure 6 where the abscissa represents the observed response, and the ordinate represents the (observed response – BLUP) estimated difference in MSE. The steepest of the three smoothed curves charts reductions in the MSE when \( k \) is known; the next steepest curve is based on BLUP estimates constructed from estimates of \( k \) and the

![Figure 4. Relative frequency distribution of per cent kcal from fat from single 24-hour intake based on 50,000 simulated days](image)
Figure 5. Relative frequency distribution of per cent kcal from fat for 25,000 simulated subjects (based on a minimum of 10 days' measures)

Figure 6. Difference in MSE (observed per cent kcal – BLUP) by observed or latent per cent kcal from fat overall mean based on duplicate measures from five subjects, applied to the same subjects. The third convex curve is based on MSE comparisons for BLUP estimates formed by applying the estimated shrinkage constants and mean to the next consecutive five subjects. The same general pattern of large reductions in MSE for BLUP estimates occurs for each BLUP estimator as when observed per cent kcal from fat measures deviate further from the mean of 31 per cent kcal. These gains are less when the mean and variance components are estimated, and even less when the estimated components are used to form estimates for the next set of subjects. For observed per cent kcal from fat measures between 29 and 33 per cent the MSE of the observed per cent kcal from fat intake was actually smaller than that of the BLUP estimate when variance components were estimated based on five subjects, with this range increasing to 26–35 per cent (representing 37 per cent of the observations) when EBLUPs were evaluated for the next five subjects based on the previous five subject’s estimates. In these ranges, however, the differences in MSE between the observed per cent kcal from fat and the BLUP estimate of per cent kcal from fat were small.

Figures 7 and 8 present histograms of the simulated total metabolic rate equivalent leisure time activity based on latent values (and response error) as reported by subjects via 24-hour recalls for
470 subjects on 6458 days, with each subject reporting a minimum of ten recall measures. A MET of one is the amount of energy used for metabolic functions while resting for one hour. The mean MET equivalent was 1.95, with subject and response error variances of 4.07 and 18.72, respectively, resulting in $k = 0.179$. The observed 24-hour recall leisure time METs were highly skewed, ranging from 0 (for 37 per cent of the simulated days) to 61.4 METs. Estimated shrinkage constants were highly variable based on study sets of $m = 5$, with negative variance component estimates occurring for 33 per cent of the study sets (resulting in estimated shrinkage constants of zero).

Subjects with latent MET values less than 5.4 METS (88.3 per cent of the subjects) had smaller average MSE based on BLUP estimates constructed from true values of $k$ compared to observed MET values (wavy concave curve in Figure 9). When comparisons were made based on observed MET values, BLUP estimates had equivalent MSE as the observed MET for observed MET values less than 4.9, after which the BLUP estimates were more accurate. The observed MET values resulted in smaller MSE estimates of the latent values when the observed MET was less than 3.2 (which occurred for 75.7 per cent of the observations), and BLUP estimates were based
on estimated variance components for groups of five subjects (applied to these subjects), and 4.9 METs (occurring for 85.5 per cent of the observations) when the variance estimates were used to form BLUP estimates for the next set of five subjects. For observed MET responses greater than these values, the BLUP estimates were more accurate, often dramatically so.

The large proportion of observations with smaller MSE for observed MET values compared to BLUP estimates stems from the sensitivity of the BLUP estimates to the estimated mean when based on small sample sizes. Since MET values are highly skewed, a single large MET value may result in pulling the observed MET values away from the subject’s true latent value towards an artificially high mean. The skewness influences the MSE comparison between BLUP and observed MET values even when parameter estimates for the BLUP values are based on larger sample sizes. With the mean and shrinkage constants estimated from $n = 25$, observed MET levels less than 3.1 METs (75 per cent of the observations) have smaller MSE than BLUP estimates; with $n = 100$, when observed MET levels are less than 2.5 (70 per cent of the observations) the observed MET levels have smaller MSE compared to estimated BLUPs.

The effect of skewness on the estimated mean can be reduced by using the sample median in place of the mean when forming the BLUP estimates (Table I). We repeated simulations constructing BLUP estimates by shrinking the observed MET response to the sample median in a similar manner. When the median and shrinkage constants were based on $n = 5$, there was little difference in the relative performance of BLUP estimates. When estimates were based on $n = 25$ or $n = 100$, the BLUP estimates had smaller MSE for substantially larger proportions of the observations.

6. DISCUSSION

When a patient walks into a physician’s office, the objective of measuring the patient’s total cholesterol (or obtaining the patient’s diet or physical activity history) is to provide information that will help in treating or counselling the patient. Interest is in the patient’s latent value for a characteristic, and not in an average based on the latent values for patients with the same measured value, a parameter known as the best predictor. The treatment/advice is offered to the specific patient who was tested, not some conceptual population of patients.
Table I. Summary range where single observed 24-hour reported MET levels have smaller MSE than BLUP estimates based on estimated shrinkage constants and estimates of the mean or median

<table>
<thead>
<tr>
<th>Group size</th>
<th>Estimator</th>
<th>Range of observed MET levels where the observed MET has smaller MSE than BLUP</th>
<th>Per cent of observed MET levels where the observed MET has smaller MSE than BLUP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Same subjects</td>
<td>Next group of subjects</td>
</tr>
<tr>
<td>5</td>
<td>Mean</td>
<td>&lt; 3:2</td>
<td>&lt; 4:9</td>
</tr>
<tr>
<td>25</td>
<td>Mean</td>
<td>&lt; 3:1</td>
<td>&lt; 3:1</td>
</tr>
<tr>
<td>100</td>
<td>Mean</td>
<td>&lt; 2:9</td>
<td>&lt; 2:9</td>
</tr>
<tr>
<td>5</td>
<td>Median</td>
<td>&lt; 3:1</td>
<td>&lt; 5:7</td>
</tr>
<tr>
<td>25</td>
<td>Median</td>
<td>0:7 to 4:8</td>
<td>0:7 to 5:6</td>
</tr>
<tr>
<td>100</td>
<td>Median</td>
<td>0:9 to 4:8</td>
<td>0:9 to 4:9</td>
</tr>
</tbody>
</table>

The basic problem faced by the provider is that for most biological and behavioural characteristics, the measured value on the patient is not equal to the patient’s latent value. How should the physician estimate the patient’s latent value? Assuming a simple additive response error model links the response to the latent value, with the latent value equal to the expected value of response, a common strategy to reduce response error is to take additional measures on the patient, and average the responses. The goal of this strategy is to replace an estimate with a parameter, or at least to reduce response error to the point where from the standpoint of medical decisions, the difference between the estimate and the parameter can essentially be ignored. As the number of measures on a patient gets larger, the patient’s average response (the BLUE) approaches the patient’s true latent value. This consistency applies equally for BLUP estimates which also approach the patient’s true latent value as \( n \) is increased.

Although this tried and true approach to reducing error works, it has its cost. The cost is in delays in time (as repeated measures are made), the expense of the measurement itself, and the cost in patient time. In addition, patients may not return for the repeated tests. Even with repeated measures, response error is never completely eliminated, leaving the nagging question of which estimator is best for the patient’s latent value. This question is typically answered by comparing properties of the estimators (that is, bias and variance (or MSE)).

The MSE can be tabulated over a sequence of single measurements on the patient (A), or over a sequence of selections of patients that would have had the observed response (B), where the response observed on the selected patient is the measured response (which is identical for all patients). Figures 3(a),(b), 6 and 9 provide a comparison of the MSE under the two conceptual sequences for a given patient’s latent value. When compared under sequence A, the BLUE estimate is not uniformly best. In fact, for the majority of the observed responses, the BLUP estimator has smaller MSE, a result that received extensive discussion over 20 years ago.\(^{11-15}\) For the examples considered, although the BLUP is not uniformly best over a range of latent values, the average MSE for the BLUP is smaller than that for the BLUE.

The MSE results favour BLUP estimates more strongly when based on sequence B. The distinction between the sequences does not result from changing the target parameter (the latent value remains the same for each MSE comparison). The distinction arises due to averaging MSEs...
using different groupings of observations. Sequence B comparisons are based on comparing the average MSE for subjects with a given observed value. For total cholesterol, the MSE for the BLUP is nearly always equivalent or smaller than that for the BLUE. We conclude that if the physician wants to estimate a patient’s latent total cholesterol based on a single measure, the BLUP estimator will be closer to the patient’s true latent value than the BLUE estimate. For per cent kcal from fat, the BLUP estimates have MSE that are similar to the MSE of BLUE estimates for observed response in the middle of the distribution, and smaller MSE elsewhere. For the highly skewed response variable of leisure time MET activity, BLUP estimates have smaller MSE only for large observed MET values.

We feel that the results are sufficiently convincing for total cholesterol and per cent kcal from fat intake to advocate the use of BLUP estimates in routine practice. Concern over the increase in MSE that will result from estimation of variance components required for BLUP estimation are mitigated by the simulation results indicating that very crude estimates (based on repeat measures on only five patients) will only modestly reduce the gain in MSE observed by BLUP estimates when the measured variable has low reliability. This result implies that only a small resource investment (in duplicate measures to estimate the variance components and the shrinkage factor) is required to tailor these results to other settings. Of course, a prudent systematic approach (such as obtaining a duplicate measure for every 100th patient seen) will practically eliminate the inflation of the BLUP MSE due to uncertainty in the mean and variance components over time.

In addition to improving the physician’s estimate of the patient’s latent value for use in determining advice or treatment, the use of BLUP estimates can have important practical impact on practice protocols and research design. If more accurate estimates can be obtained from the same data by simply using a different estimator, then the cost of collecting additional data to improve such accuracy (using a BLUE estimator) can be saved. For example, the Seasons study was designed to estimate dietary intake and physical activity on study subjects in a three-month period. Since day-to-day variability in these response variables was expected to be high, the study was designed to collect three 24-hour recall measures, and then use the average of these measures to estimate the subject’s quarterly intake or activity. Table II compares the MSE of the average of these three 24-hour measures (the BLUE), with the MSE of a BLUP estimate using a single 24-hour measure. Although the MSEs are comparable (indicating equivalent accuracy of the subject’s latent values), only one-third of the collection cost is required to form the BLUP estimates.

The BLUP estimates are effectively empirical Bayes estimates, and can be motivated from a Bayesian perspective. This motivation has not been emphasized since we concentrate on estimating latent values, not linear predictors. The connection is important, however, since it points to strategies used in a Bayesian framework (such as setting a prior distribution on variance components) that may lead to improved BLUP estimates when the shrinkage factor is based on small sample sizes.

It is important to note considerations that limit generalization of these results to other biologic and patient factors. These considerations relate directly to the distribution of latent values for the subjects. BLUP estimates are not best predictors, and do not provide unbiased estimates of the best predictor unless the latent values are normally distributed. The shape of the latent distribution will influence the properties of the BLUP estimates relative to the BLUE estimate. For unimodal distributions, since the BLUP estimator systematically shrinks the observed response towards the mean, the MSE of the BLUP estimators will often be reduced relative to the BLUE estimate. For heavy tailed skewed distributions, as illustrated by the example of leisure time
Table II. Comparison of MSE for observed average and BLUP estimates of per cent kcal from fat based on study designs with one measure (1X) and three measures (3X) per subject from the Seasons study (493 subjects)

<table>
<thead>
<tr>
<th>Per cent of kcal from fat intake (1X)</th>
<th>Per cent of subjects (1X)</th>
<th>Per cent of subjects (3X)</th>
<th>Average MSE observed (3X)</th>
<th>Average MSE BLUP (1X)</th>
<th>Average MSE BLUP (3X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–10</td>
<td>1.9</td>
<td>0.5</td>
<td>99.8</td>
<td>29.2</td>
<td>16.6</td>
</tr>
<tr>
<td>10–20</td>
<td>11.8</td>
<td>8.1</td>
<td>42.6</td>
<td>24.9</td>
<td>17.5</td>
</tr>
<tr>
<td>20–30</td>
<td>34.5</td>
<td>40.3</td>
<td>16.3</td>
<td>21.3</td>
<td>12.4</td>
</tr>
<tr>
<td>30–40</td>
<td>36.6</td>
<td>42.1</td>
<td>15.2</td>
<td>19.5</td>
<td>11.7</td>
</tr>
<tr>
<td>40–50</td>
<td>13.2</td>
<td>8.5</td>
<td>41.1</td>
<td>21.2</td>
<td>16.0</td>
</tr>
<tr>
<td>50–60</td>
<td>1.8</td>
<td>0.5</td>
<td>101.6</td>
<td>25.3</td>
<td>17.2</td>
</tr>
<tr>
<td>60–70</td>
<td>0.1</td>
<td>0.01</td>
<td>221.2</td>
<td>40.5</td>
<td>14.6</td>
</tr>
<tr>
<td>Overall average</td>
<td>20.94</td>
<td>21.29</td>
<td>12.87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MET, shrinkage towards the median will reduce the MSE for BLUP estimators, but the reduced MSE may be limited to observed values in the tails of the distribution. For the patient factors considered in this study, the latent value distributions were unimodal. In many other settings for these same patient factors, one might expect a similar unimodal distribution, with modest differences in parameters, and hence superiority of BLUP estimates.

For factors that have latent distributions that are not unimodal, BLUP estimators may be less accurate than BLUE estimates. Settings where the latent distribution follows a U-shape will perform particularly poorly. In addition, when the latent distribution consists of a finite population and latent values in the tails are widely separated, the BLUP estimates may perform poorly for observed responses in these tails even if the overall distribution is unimodal. Such a situation was evident for a few patients for cholesterol in the simulation studies (Figure 3). More research is called for to further clarify and generalize when the BLUP estimates may fail. However, the simulation results clearly demonstrate that the BLUE estimates fail much more commonly relative to BLUP when accuracy is compared. These results lead us to the conclusions that for variables like total cholesterol and per cent kcal from fat, BLUP estimates should be the estimator of choice. Why not use BLUPs?

ACKNOWLEDGEMENT

This work was supported by the NHLBI:R01-HL52745.

REFERENCES


