STATISTICS IN MEDICINE
Statist. Med. **19**, 83–98 (2000)

PATIENT-SPECIFIC ANALYSIS OF SEQUENTIAL HAEMATOLOGICAL DATA BY MULTIPLE LINEAR REGRESSION AND MIXTURE DISTRIBUTION MODELLING

C. E. McLAREN[1*], E. L. KAMBOUR[2], G. J. McLACHLAN[3], H. C. LUKASKI[4], X. LI[5], G. M. BRITTENHAM[6] AND G. D. McLAREN[7,8]

[1] *Division of Epidemiology, Department of Medicine and Chao Family Comprehensive Cancer Center, University of California, Irvine, CA 92697, U.S.A.*

[2] *Department of Statistics, Texas A&M University, College Station, TX 77840, U.S.A.*

[3] *Department of Mathematics, The University of Queensland, Brisbane, Qld 4072, Australia*

[4] *USDA, ARS, Grand Forks Human Nutrition Research Center, Grand Forks, ND 58202, U.S.A.*

[5] *University of North Dakota School of Medicine and Department of Veterans Affairs Medical Center, Fargo, ND 58102, U.S.A.*

[6] *Columbia University College of Physicians and Surgeons, New York, NY 10032, U.S.A.*

[7] *Division of Hematology/Oncology, Department of Medicine and Chao Family Comprehensive Cancer Center, University of California, Irvine, CA, U.S.A.*

[8] *Department of Veterans Affairs Medical Center, Long Beach, CA 90822, U.S.A.*

SUMMARY

Automated storage and analysis of the results of serial haematologic studies are now technically feasible with present-day laboratory instruments and devices for data storage and processing. In current practice, physicians mentally compare a laboratory result with previous values and use their clinical judgement to determine the significance of any change. To provide a statistical basis for this process, we describe a new approach for the detection of changes in patient-specific sequential measurements of standard haematologic laboratory tests. These methods include hierarchical multiple regression modelling, with a weighted minimum risk criteria for model selection, to choose models indicating changes in mean values over time. This study is the first to analyse sequential patient-specific distributions of laboratory measurements, utilizing mixture distribution modelling with systematic selection of starting values for the EM algorithm. To evaluate these statistical methods under controlled conditions, we studied 11 healthy human volunteers who were depleted of iron by serial phlebotomy to iron-deficiency anaemia, then treated with oral iron supplements to replete iron stores and correct the anaemia. Application of sequential patient-specific analyses of haemoglobin, haematocrit, and mean cell volume showed that significant departures from past values could be identified, in many cases, even when values were still within the population reference ranges. Additionally, for all subjects sequential alterations in red blood cell volume distributions during development of iron-deficiency anaemia could be characterized and

* Correspondence to: Christine E. McLaren, Division of Epidemiology, School of Medicine, University of California, Irvine, 224 Irvine Hall, Irvine, CA 92697-7550, U.S.A. E-mail: cmclaren@uci.edu

Contract/grant sponsor: National Institutes of Health
Contract/grant number: R15 HL48349
Contract/grant sponsor: Fogarty International Center
Contract/grant number: FO6TWO2117
Contract/grant sponsor: University of Queensland

CCC 0277–6715/2000/010083–16$17.50

Received August 1996
Accepted February 1999

quantified. These methods promise to provide more sensitive techniques for improved diagnostic evaluation of developing anaemia and serial monitoring of response to therapy. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Currently physicians compare individual blood test values with population-based reference ranges to assist in the evaluation and management of common haematologic (blood) disorders such as anaemia. There is considerable evidence, however, that for many blood constituents, the average amount of within-person variation over time (excluding measurement error) is much less than the between-person variation.[1–3] Thus, when analytic variation is reduced, the magnitude of variation in haematologic parameters measured in healthy individuals over time will be small in comparison to the reference ranges in use. An observation that is within the population-based reference range may represent a clinically significant deviation from one subject's usual condition, whereas another result, outside population-based limits, may just represent expected random variation for another subject.

With these considerations in mind, a subject's own prior laboratory record may be a better guide to proper assessment of current findings in monitoring abnormalities or the effects of therapy. Ross *et al.*[4] examined intra-individual versus inter-individual variation (variation among individuals) of haematologic parameters in healthy individuals over a period of nine years. They demonstrated that the haematologic parameters measured as part of an automated complete blood count (CBC) and differential are quite stable despite instrumentation changes during the study. They concluded that, for some parameters, comparison with reference values derived from previous data would be a more sensitive detector of abnormality than comparison of a single value with a normal population range. Examination of red blood cell volume distributions has shown that for distributions measured using aperture impedance technology, variability between samples drawn from different individuals exceeded variability between samples measured from the same individual.[5, 6] This evidence suggests the need for statistical methods that provide comparison to reference values specific to the individual for sequential monitoring.

While computer technology is now available for the storage and analysis of vast amounts of data from patients, to date no statistical methods for evaluation of serial haematologic measurements have been established for the evaluation of patients in situations with anaemia when few, if any, initial steady-state observations may be available for an individual patient. The specific hypothesis underlying this project was that patient-specific examination of serial haematologic measurements would provide a sensitive method for the early detection and diagnosis of developing iron-deficiency anaemia. This provided rationale for the development of statistical techniques for sequential analysis of longitudinal data collected from a single individual. For this purpose we analysed blood test values collected over time by hierarchical multiple regression modelling to estimate changes in the mean response. We also discuss methods for sequential analysis of serial red blood cell volume distributions. By application of these methods to data obtained from volunteers during iron depletion, we demonstrate detection of patient-specific significant changes in haematologic measurements.

Although a variety of theoretical approaches have been published for analysis of shifts in the mean of longitudinal data values, some methods were designed to detect only a single shift in the mean response[7–10] and others lack specific application to current problems in medicine and applied sciences.[11, 12] In this paper, we demonstrate by application to real data the advantages

of the statistical procedures that we describe. For example, we illustrate that changes can be identified during the development of iron-deficiency anaemia even when values are still within population reference ranges. We also describe a valid clinical application of sequential analysis of red blood cell volume distributions. This study is the first to use mixture distribution modelling with systematic selection of starting values for the EM algorithm to quantify sequential changes in red blood cell volume distributions during the development of iron-deficiency anaemia. On a daily basis, physicians face the problem of analysing an accumulated wealth of information on individual patients and they must make clinical decisions based upon comparison of serial laboratory test values. We suggest that the methods we describe provide effective statistical tools for this clinical decision-making process. A major advantage is that the techniques could readily be applied in the clinical setting to provide a statistical basis for computerized review of laboratory data to assist in diagnosis of disease and monitoring of response to therapy.

2. METHODS

2.1. Multiple regression modelling with longitudinal data

Our major goal was to develop statistical methods to detect sequential changes in the mean response for laboratory blood tests. Consider the multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{n-1,i} + \varepsilon_i \tag{1}$$

where ε_i is i.i.d. $N(0, \sigma^2)$ $(i = 1, \ldots, n)$. Here σ^2 represents the true variance in the response Y about the regression and is assumed to be the same for any given blood sample if the model is correct. For n sequential blood samples collected from the same individual let Y_i represent the response for sample i and let $X_{1i}, X_{2i}, \ldots, X_{n-1,i}$ be indicator variables where $X_{pi} = 0$ for $i \leqslant p$ and $X_{pi} = 1$ for $i > p$ $(p = 1, \ldots, n-1)$. Stating the model as

$$Y = X\beta + \varepsilon \tag{2}$$

where the first column of matrix X consists of n 1's, the least square estimator of β is then given by $b = (X'X)^{-1}X'Y$, where $b = (b_0, b_1, \ldots, b_p)$.

For subsets of indicator variables, multiple regression models were formed in a hierarchial stepdown fashion. All possible subsets of k indicator variables, $(k = n-2, n-3, \ldots, 1)$, were considered using the method known as leaps and bounds.[13] A restriction on the analysis was that the full model with $p = n-1$ indicator variables was never considered since the degrees of freedom for error for the regression MSE would be zero, where

$$\text{MSE} = (Y'Y - b'X'Y)/(n - [p+1]). \tag{3}$$

This statistical restriction does not preclude application of the technique because in laboratory medicine it is a rare event to observe a medically meaningful change at every sequential observation of a laboratory blood test for a single individual.

2.1.1. Minimum Risk

Criteria such as R^2, or the adjusted R^2, when used as model selection criteria will be biased toward larger models, thus our criterion for final model selection was based on the minimum *risk* (that

is, expected *loss*). The loss is a function of the (squared) bias and variance of the estimator of the mean response. As shown by Eubanks (1988),[14] an unbiased estimator of risk, is

$$\hat{P} = \frac{\text{SSE}}{n} + \frac{2\sigma^2 k}{n}. \tag{4}$$

Eubanks notes that $\hat{P}$ is closely related to the general criterion for model selection, Mallows C_p. One possible estimator for σ^2 would be the regression MSE $= \frac{\text{SSE}}{n}$. However, this estimator will be different for each model, while the variance of Y_i is assumed to be the same for each model. It would be preferable to have an estimator of σ^2 that would not be model dependent. A strongly consistent estimator of σ^2 that does not depend on the model is that proposed by Gasser, Sroka and Jenner-Steinmetz[15] σ^2_{GSJS} where

$$\hat{\sigma}^2_{\text{GSJS}} = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \tilde{\varepsilon_i}^2 \tag{5}$$

and

$$\tilde{\varepsilon_i} = Y_i - \frac{Y_{i-1} + Y_{i+1}}{2}. \tag{6}$$

For our modelling, we used $\hat{\sigma}^2_{\text{GSJS}}$ to estimate the true variance of the response Y, about the regression.

2.1.2. Model selection

The estimated risk was computed for each model selected by the leaps and bounds method and compared to that of the null model (no changes in the mean response). Initially, we considered selecting the 'optimal' model, as the model with the minimum unbiased risk. However, by simulation, we found that the rate of rejection under the null model was high and increased rapidly as the length of the sequence increased. Thus to reduce the type I error rate, we multiplied the unbiased risk for each model, selected by the leaps and bounds method, by a weight λ raised to the power k, where k was the number of model indicator variables, that is, potential changes in the mean response. As motivation for our choice of the weight λ, assume that there is a constant probability, p, that a patient will have a change in a given blood test value. The probability associated with a particular model with n blood test values and k changes in the mean response, will be $(p^k)(1-p)^{n-k}$. When the number of changes considered is increased by one (for example, considering a model with two changes in the mean response compared to a model with three changes) then this probability decreases by a factor of $p/(1-p)$ and when a model with no changes is compared to a model with k changes, this probability decreases by a factor of $(p/(1-p))^k$. Thus to take a conservative approach to model selection, we multiplied the estimated unbiased risk for a given model by λ^k with λ^k representing $((1-p)/p)^k$.

The estimated weights λ were found using simulations with an iterative search for the sample-size dependent value of λ that corresponded to a specified type I error rate. Note that when models with the same sequence length were compared, those with more predictors would be penalized. Values for λ are given Table I for sequences of length 3 to 15, typical sequence lengths for our studies. For example, for a sequence of length 5 and a desired significance level of 0·05, the estimated risk for the one-change model selected by the leaps and bounds method was multiplied by 3·68.[1] Similarly, the estimated risk for the two-change model selected by the leaps and bounds

Table I. Values of the weight λ for selected empirical significance levels

Sequence length	Significance level			
	0·25	0·10	0·05	0·01
3	1·57	5·14	7·92	8·50
4	1·48	3·78	5·80	8·18
5	1·39	2·41	3·68	7·85
6	1·36	2·17	3·14	6·27
7	1·33	1·93	2·60	4·68
8	1·32	1·83	2·38	4·18
9	1·30	1·72	2·16	3·61
10	1·29	1·67	2·05	3·29
11	1·27	1·62	1·94	2·97
12	1·26	1·57	1·85	2·70
13	1·24	1·51	1·76	2·43
14	1·23	1·48	1·71	2·35
15	1·22	1·45	1·66	2·26

method was multiplied by $3{\cdot}68^2$. The weighted risk estimates were then compared to the estimated risk for the null (no-change) model and the 'optimal' model was chosen as the one with the smallest weighted estimated risk.

2.2. Model diagnostics

The regression methods for detecting sequential changes in the mean assume independent errors over time, normality and constant variance on the error structure. Since in haematologic studies, typically few observations would be available (for example, 10 to 12 at most), these assumptions are difficult to test rigorously. However, to check the fit of the selected optimal model and to look for patterns in the residuals, we examined plots of the residuals over time, plots of the residuals versus the fitted values, and normal probability plots of the residuals and the Studentized residuals. To take into account the autocorrelation in the time-series data, transformations of the response vector **TY** and the corresponding predictor matrix **TX** were made and the hierarchical regression method was applied. Note that the resulting regression coefficients from analysis of the transformed data, represent the same characteristics as those in the untransformed model, since the same linear transformation was applied to both the predictors and the response.

The correlation matrix, **R**, for a sequence of first-order autocorrelated data is

$$\mathbf{R} = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \ddots & \vdots \\ \rho^2 & \rho & 1 & \ddots & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^{n-1} & \dots & \rho^2 & \rho & 1 \end{pmatrix}.$$

Let Λ be the Cholesky root of $\mathbf{R}$, that is, $\Lambda\Lambda' = \mathbf{R}$, and let $\mathbf{T} = \Lambda^{-1}$. $\mathbf{T}$ is the transformation matrix such that the elements of $\mathbf{TY}$ will be uncorrelated. That is

$$\text{var}(\mathbf{TY}) = \mathbf{T}\text{var}(\mathbf{Y})\mathbf{T}' = \mathbf{T}\sigma^2\mathbf{R}\mathbf{T}' = \sigma^2\Lambda^{-1}\Lambda\Lambda'(\Lambda^{-1})' = \sigma^2\mathbf{I}.$$

Since ρ is unknown, it is estimated using the sample first-order autocorrelation, r.

2.3. Haematologic methods

In order to evaluate these methods under controlled conditions, we studied seven female and four male healthy human volunteer subjects who were depleted of iron by serial phlebotomy to iron-deficiency anaemia, then treated with oral iron supplements to replete iron stores and correct the anaemia. The study was conducted with the approval of the University of North Dakota Institutional Review Board and the U.S. Department of Agriculture Human Study Committee. Written, informed consent was obtained from each volunteer before participation in any testing. At weekly intervals during the study, a blood sample of 5·0 ml was drawn from each volunteer and two types of data were collected: (i) serial measurements of individual laboratory tests including haemoglobin (Hb), haematocrit (Hct), and mean cell volume (MCV); (ii) serial red blood cell volume distributions. Complete blood count values were determined using a Coulter particle counter model S-Plus VI (Coulter Electronics, Inc., Hialeah, Fla.). Body iron status was evaluated by serum ferritin. From each volunteer, an average volume ($\pm$1 SD) of 1455 ± 883 ml of blood was removed over 36 $\pm$ 18·6 days, accompanied by a decrease in haemoglobin from 14·3 $\pm$ 1·2 g/dl to 11·5 $\pm$ 1·2 g/dl and a post-phlebotomy serum ferritin of 6·9 $\pm$ 2·3 μg/l.

Blood samples (5·0 ml) were also drawn from 11 additional healthy adult volunteers (four females and seven males) and serial measurements of the haematocrit were performed over a two-week period. These individuals were not depleted of iron by serial phlebotomy, thus they formed a reference sample group for the analysis of serial measurements of Hct under haematologically normal conditions. The study was conducted using a protocol approved by the Institutional Review Boards of the University of North Dakota and the Veterans Affairs Medical Center. Reference ranges for the laboratory were as follows: Hb $13{\cdot}5 - 17{\cdot}5$ g/dl (males), 12–16 g/dl (females); Hct 42–52 (per cent) (males), 37–47 (per cent) (females); MCV 80–100 fl.

2.4. Simulation

An increase or decrease of at least 1 gm/dl in haemoglobin is considered to have clinical significance. This represents a change of 2·0 standard deviations in the mean haemoglobin.[16] Considering three scenarios, we designed a simulation to evaluate the conditions under which sequential changes in the mean haemoglobin could be detected with the proposed statistical procedure. For each scenario, 10,000 sequences of length $3, 5, 7, \ldots, 15$ were generated. The simulated scenarios were as follows: (i) a shift in the mean of 1 standard deviation occurred at the middle value in the sequence. For each sequence, values were generated in which all data prior to the middle value were N(0,1) and data at or after the middle value were N(1, 1); (ii) a shift in the mean of 2 standard deviations occurred at the middle value in the sequence. Values were generated in which all data prior to the middle value were N(0,1) and data at or after the middle value were N(2, 1); (iii) a shift in the mean of 3 standard deviations occurred at the middle value in the sequence. Values were generated in which all data prior to the middle value were N(0,1) and data at or after the middle value were N(3, 1).

For each sequence, we recorded the number of simulations (out of 10,000) in which a shift in the mean was detected using the weighted unbiased risk approach at significance levels of $0{\cdot}25, 0{\cdot}10, 0{\cdot}05$ and $0{\cdot}01$.

2.5. Red blood cell volume distributions

Automated haematology analysers now determine the volume of red blood cells and routinely provide the distribution of red blood cell volumes. To promote methodological uniformity in cell size studies, the Expert Panel on Cell Cytometry of the International Council for Standardization in Haematology (ICSH) has made recommendations of general principles for the analysis of cell volume data.[17, 18] These recommendations include fitting an observed cell volume distribution to a reference log-normal distribution using the expectation-maximization (EM) algorithm for parameter estimation as described by McLaren and colleagues.[5, 19] Distributions showing a poor fit to a single log-normal model should be examined for the presence of more than one population of cells.

To study patient-specific sequential changes in red blood cell volume distributions during iron depletion, red blood cell volume determinations were performed for each blood sample using a Coulter particle counter model S-Plus VI as described by McLaren *et al.*[19] For each sample, frequency counts for 256 cell volume intervals of length 1·3125 fl with range 24 to 360 fl were measured in duplicate. Artefactual frequency counts occur in the upper ranges of cell volume for particle counts resulting from cell coincidence, doublets, triplets, and agglutinated cells, and, in the lower ranges, for counts resulting from platelet clumps, large platelets, and electrical interference. To eliminate artefactual frequency counts, each distribution was doubly truncated using a truncation algorithm developed specifically for red blood cell volume distributions.[5]

2.6. Parameter estimation and starting values for the EM algorithm

After truncation, each distribution was tested for best fit to a single log-normal distribution or a mixture of two log-normal distributions. McLachlan and Basford[20] and McLaren *et al.*[21] discuss iterative computation of the maximum likelihood estimates for mixture models via the EM algorithm of Dempster *et al.*[22] To avoid problems with convergence of the EM algorithm due to poor choice of starting values or multiple roots of the likelihood equation, a systematic procedure permitting selection of multiple starting values for the parameters was devised.[23] A subroutine provided by Jones and McLachlan[24] for fitting by maximum likelihood a g-component normal mixture was modified to fit a mixture of g log-normal components for $g = 1$ and $g = 2$ in turn, without need to specify the starting value for the vector of unknown parameters.

2.7. Likelihood ratio test, resampling, and goodness-of-fit

A resampling approach was applied to assess the P-value of the likelihood ratio statistic $-2\log\lambda$ for the test of $g = 1$ versus $g = 2$ components in the log-normal mixture.[25] For a significance test with approximate size $\alpha = 0{\cdot}05$, 1000 independent bootstrap samples were generated under the null hypothesis of a single log-normal distribution. Then replications of $-2\log\lambda$ were obtained. The null hypothesis of a single log-normal model was rejected if the value of $-2\log\lambda$ obtained from the real data was greater than the 950th largest of the bootstrap replications.

Table II. Per cent of samples out of 10,000 in which a shift in the mean of Δ standard deviations was detected at specified empirical significance levels

Δ	Sequence length	Significance level			
		0·25	0·10	0·05	0·01
1	3	27·9	11·6	5·7	1·5
	5	35·4	15·9	8·4	2·1
	7	40·8	20·9	11·4	3·2
	9	46·7	26·4	15·6	4·5
	11	51·0	29·4	18·6	5·4
	13	54·3	33·3	22·0	7·8
	15	58·9	38·2	25·0	8·8
2	3	32·6	12·6	6·2	1·5
	5	54·3	27·8	16·9	5·0
	7	69·3	44·6	29·0	10·5
	9	80·0	59·0	42·4	16·6
	11	86·9	69·1	53·6	24·2
	13	91·5	78·6	65·2	37·5
	15	94·3	84·8	74·3	51·4
3	3	34·5	12·6	6·2	2·5
	5	70·0	36·9	22·9	7·4
	7	88·7	66·5	46·4	19·8
	9	96·0	84·9	70·0	34·9
	11	98·8	93·5	85·0	53·9
	13	99·4	97·4	93·4	75·3
	15	99·9	98·9	96·9	85·9

3. RESULTS

3.1. Simulation

Simulation results are shown in Table II. For models with a single change in the mean from $N(0,1)$, as the sequence length increased, the proportion of sequences in which a change was detected also increased. As expected, there was low power for detection of a shift in the mean of 1 standard deviation. However, the power improved for detection of a shift of 2 standard deviations. For example, with 15 observations, the change was detected 85 per cent of the time at the 10 per cent level, and 74 per cent of the time at the 5 per cent significance level. With as few as 11 values, shifts of 3·0 standard deviations in the mean were detected for 85 per cent of the simulated sequences at the 5 per cent significance level.

3.2. Multiple regression modelling with longitudinal data

The haematological data from a representative female subject are shown in Table III. A total of 10 blood samples were obtained from this subject, over an 11 week period, at which time a state of iron-deficiency anaemia was established and oral iron replacement therapy was begun. The 10 test days are given as $-77, -62, \ldots, 0$, representing days before iron repletion therapy began. Graphical representations of the results of two separate analyses of sequential laboratory test values are shown in Figures 1 and 2. The test day is shown on the horizontal axis and corresponding

Table III. Haematological data for a representative female subject

Sample	Test day	Hb	Hct	MCV
1	−77	13·2	38·3	89·4
2	−62	12·2	36·7	89·2
3	−56	12·7	38·7	89·5
4	−49	11·4	34·3	88·9
5	−41	11·4	34·6	89·0
6	−35	11·0	33·3	89·5
7	−28	11·7	35·5	88·7
8	−21	11·0	34·0	87·4
9	−14	10·2	30·7	86·4
10	0	10·7	33·0	84·1

laboratory test values are shown on the vertical axis. The solid arrows show test days at which changes in mean test values were detected. The tip of each arrow indicates the mean laboratory test value estimated using multiple regression modelling.

3.2.1. Haemoglobin and mean cell volume

In Figure 1(*a*), analysis of the first seven haemoglobin values, from test days $-77, -62, -56, \ldots, -28$, shows that by the fourth week (test day -49) the mean haemoglobin had decreased significantly from an estimated mean of 12·6 g/dl (95 per cent confidence interval: 12·4 g/dl, 12·9 g/dl) to an estimated mean of 11·3 g/dl (11·0 g/dl, 11·7 g/dl) at test day -49. The raw unbiased risk estimate was $\hat{P} = 0{\cdot}426$ compared to $\hat{P} = 1{\cdot}248$ for the null model. For empirical significance levels of 0·05 and 0·01, the corresponding weighted unbiased risks were 1·11 and 1·99, respectively, giving an empirical p-value for the procedure of $0{\cdot}05 < p < 0{\cdot}01$. The individual t-statistic for this model was $t = -7{\cdot}84$, d.f. $= 5$, ($p = 0{\cdot}0005$) for the regression parameter representing test day -49.

From the final analysis of 10 data points, a further significant decrease in the mean was observed at the ninth week (test day -14; Figure 1(b)). The best regression model with adjustment for first-order autocorrelation between values, identified by the weighted minimum risk, considering data from 10 test days, confirmed a decrease from the initial estimated mean of 12·6 g/dl (12·4 g/dl, 12·9 g/dl) to an estimated mean of 11·7 g/dl (11·3 g/dl, 12·1 g/dl) at test day -49, with a further decrease to an estimated mean of 10·5 g/dl (9·8 g/dl, 11·0 g/dl) at test day -14. The raw unbiased risk estimate was $\hat{P} = 0{\cdot}399$ compared to $\hat{P} = 1{\cdot}785$ for the null model. For empirical significance levels of 0·05 and 0·01, the corresponding weighted unbiased risks were 1·676 and 4·318, respectively, giving an empirical p-value for the procedure of $0{\cdot}05 < p < 0{\cdot}01$. The individual t-statistics for this model were $t = -4{\cdot}74$, d.f. $= 7$, ($p = 0{\cdot}002$) for the regression parameter representing test day -49 and $t = -8{\cdot}45$ ($p = 0{\cdot}0001$), for the regression parameter representing test day -14. A clinical interpretation of the coefficient of determination of 99·9 per cent would be that with this model, after taking into account the autocorrelation between the values, virtually all of the variation in serial haemoglobin values was due to development of iron deficiency.

As shown in Table III and in Figure 2, the MCV for this subject remained within the reference range throughout the study, however, changes in the mean were identified statistically. Analysis

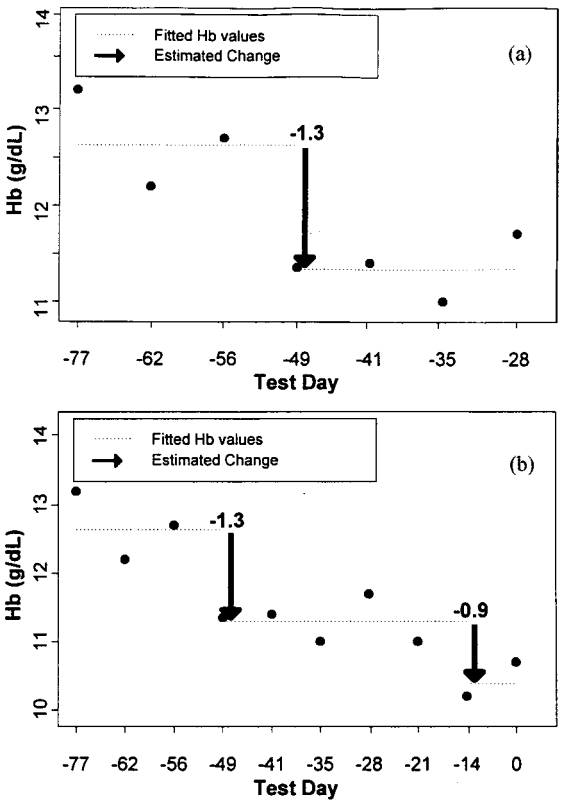

Figure 1. Sequential multiple regression analyses for haemoglobin. Arrows indicate a significant decrease in the mean haemoglobin value. Dashed lines indicate the current estimated mean haemoglobin based upon significant indicator variables for the regression model. For purposes of illustration the distance between sequential test days are shown as being equal although the actual times between sequential test days may differ

of data at the eighth week showed an initial decrease from an estimated mean of 89·2 fl (89·0 fl, 89·4 fl) to 87·3 fl (86·7 fl, 87·9 fl) at test day −21, still well within the reference range for healthy individuals. The unbiased risk estimate was $\hat{P} = 0{\cdot}167$ compared to $\hat{P} = 0{\cdot}532$ for the null model. For empirical significance levels of 0·05 and 0·01, the corresponding weighted unbiased risks were 2·38 and 4·18, respectively, giving an empirical p-value for the procedure of $0{\cdot}05 < p < 0{\cdot}01$. The individual t-statistic for this model was $t = -6{\cdot}15$, d.f. = 6, ($p = 0{\cdot}0008$), for the regression parameter representing test day −21.

Further decreases were detected by the tenth week of the study, although the final observed and estimated mean values were still within the reference ranges for females (Table III, Figure 2(b)). The best regression model identified by the weighted minimum risk criteria, confirmed a significant decrease from an estimated mean of 89·2 fl (89·0 fl, 89·4 fl) to 86·9 fl (86·4 fl, 87·3 fl) at test day −21, and a further decrease to 84·0 (83·2 fl, 84·8 fl), still well within the reference range for healthy individuals. The unbiased risk estimate was $\hat{P} = 0{\cdot}219$ compared to $\hat{P} = 3{\cdot}807$ for the null model. For empirical significance levels of 0·01, the corresponding weighted unbiased risks was 2·37 with an empirical p-value for the procedure of $p < 0{\cdot}01$ The individual t-statistics for this

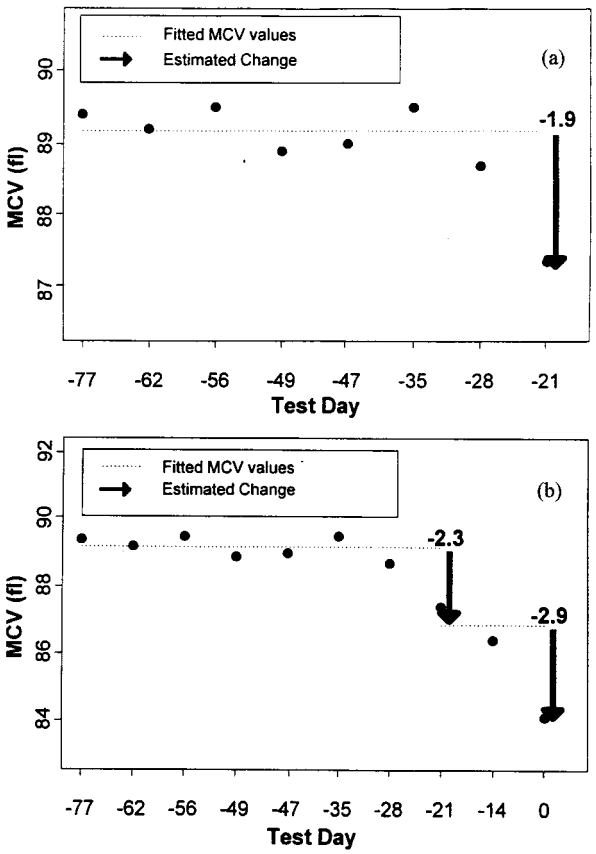

Figure 2. Sequential multiple regression analyses for mean corpuscular volume. Arrows indicate a significant decrease in the mean MCV value. Dashed lines indicate the estimated mean MCV based upon significant indicator variables for the regression model. For purposes of illustration the distance between sequential test days are shown as being equal although the actual times between sequential test days may differ

model were $t = -9{\cdot}76$, d.f. $= 7$, ($p < 0{\cdot}0001$), for the regression parameter representing test day -21 and $t = -6{\cdot}71$, d.f. $= 7$, ($p = 0{\cdot}0003$) for the regression parameter representing test day 0.

3.2.2. Overall assessment

To make an overall assessment haematologically, we examined the final analysis of each of the three laboratory tests (Hb, Hct, MCV) for the 11 volunteers. On the basis of the weighted unbiased risk, significant changes were observed at the 0·05 per cent significance level in Hb, Hct, or MCV for 9 of 11 subjects.

Analysis of Hb: the distribution of empirical p-values for the best fitting models was $0{\cdot}10 < p < 0{\cdot}25$ ($n = 2$), $0{\cdot}05 < p < 0{\cdot}10$ ($n = 2$), and $0{\cdot}01 < p < 0{\cdot}05$ ($n = 7$). The median number of changes observed at a significance level of 0·05 was 2, with an average magnitude of 1·5 g/dl or approximately 3 standard deviations in haemoglobin. In this subgroup of 9 volunteers, the minimum

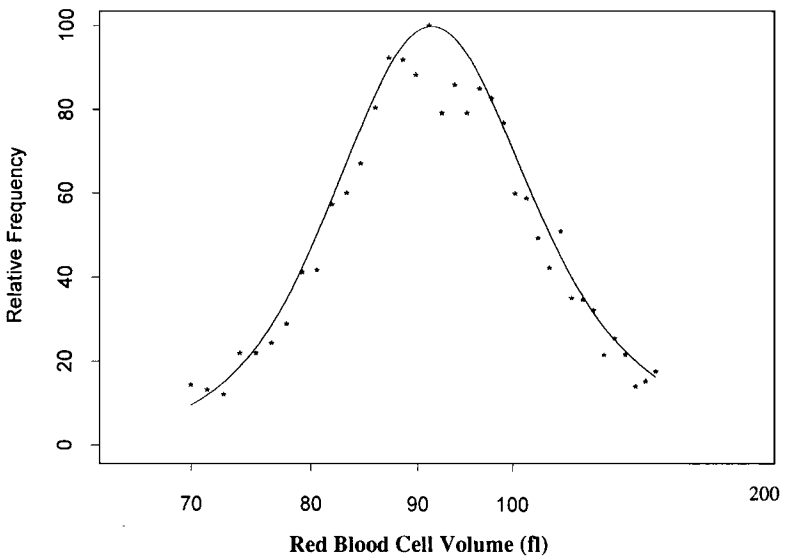

Figure 3. Distribution of red blood cell volumes plotted on a natural logarithm scale. Solid lines represent the fitted mixture distribution for two log-normal subpopulations. The relative frequency is calculated as 100 per cent (frequency count/maximum frequency count)

and maximum estimated decreases in mean haemoglobin were 0·6 fl and 3·5 g/dl, respectively. For 2 of these 9 subjects (22 per cent), changes were identified in Hb even though the values were still within the population reference ranges. Thus the sequential analysis procedure provided comparison of individual values to reference values derived from previous data and gave a sensitive detector of abnormality. The information gained is more informative than that of comparison to a normal population range. The fit of separate models was also assessed by examining the residuals, that is, the difference between each observed haemoglobin value and the corresponding predicted mean value. For example, residuals were examined for the five (independent) comparable models in which two sequential changes in the mean haemoglobin concentration were detected. The sample size (sequence of haemoglobin values) from an individual subject was small for a full residual analysis, but residuals plotted against predicted mean haemoglobin concentration were evenly distributed around zero without discernible patterns.

Analysis of Hct: the distribution of empirical p-values for the best fitting models was $0{\cdot}10 < p < 0{\cdot}25$ $(n = 4)$, $0{\cdot}05 < p < 0{\cdot}10$ $(n = 1)$, $0{\cdot}01 < p < 0{\cdot}05$ $(n = 2)$, and $p < 0{\cdot}01$ $(n = 4)$. The median number of changes observed for a significance level of 0·05 was 1, with an average magnitude of 4·5 per cent. The minimum and maximum estimated decreases in haematocrit for this subgroup of 6 volunteers were 1·8 per cent and 5·7 per cent, respectively. For 2 of these 6 subjects (33 per cent), changes were identified in Hct even though the values were still within the population reference ranges.

Analysis of MCV: we examined models of weekly sequential changes in average MCV for all subjects, considering data from the first to the last study week, where from 5 to 12 observations were analysed. The distribution of empirical p-values for the best fitting models was $p > 0{\cdot}25$ $(n = 2)$, $0{\cdot}1 < p < 0{\cdot}25$ $(n = 1)$, $0{\cdot}05 < p < 0{\cdot}10$ $(n = 3)$, $0{\cdot}01 < p < 0{\cdot}05$ $(n = 4)$, and $p <$

0·01 ($n = 1$). Of the five subjects with significant changes in the mean MCV, using a significance level of 0·05, the average absolute magnitude of change in the mean MCV was 3·05 fl, with minimum and maximum of 1·5 fl and 5·3 fl. For each of these 5 subjects changes were identified while values for MCV remained within the population reference ranges, clearly indicating a subject-specific response to the iron depletion. No discernible patterns were found for residual analyses of the three comparable models in which two sequential changes in the average MCV were detected.

These procedures were also applied to data from a reference sample group of 11 healthy adult volunteers to study the analysis of serial measurements of haematocrit under haematologically normal conditions. We tested baseline and two subsequent weekly values for each subject. As expected, for each volunteer no changes were observed in Hct over time with an empirically determined p-value $> 0{\cdot}25$ for all. In contrast, the distribution of empirical p-values for the analysis of values drawn on the first three test days from volunteers depleted of iron was $p > 0{\cdot}25$ ($n = 3$), $0{\cdot}1 < p < 0{\cdot}25$ ($n = 5$), $0{\cdot}05 < p < 0{\cdot}10$ ($n = 1$), $0{\cdot}01 < p < 0{\cdot}05$ ($n = 2$).

3.3. Red blood cell volume distributions

Sequential changes in red blood cell volume distribution were observed during the study period in all subjects. For example, for the subject whose individual haematological test results are shown in Table III and discussed above, the red blood cell volume distribution from the first test day (-77) was observed to fit a single log-normal distribution quite well (likelihood ratio statistic: $-2 \log \lambda = 0{\cdot}73$, bootstrap $p = 0{\cdot}937$; goodness-of-fit $\chi^2 = 33{\cdot}6$, d.f. $= 35$, $p = 0{\cdot}54$) with $e^{\mu} = 94{\cdot}6$, which is within the reference range for healthy individuals reported by McLaren *et al.* At the 5th week of the study, that is, test day -41, the distribution of red blood cell volumes was altered in response to phlebotomies and could be seen to fit two log-normal distributions (Figure 3; likelihood ratio statistic $-2 \log \lambda = 11{\cdot}59$, bootstrap $p = 0{\cdot}012$; goodness-of-fit $\chi^2 = 43{\cdot}7$, d.f. $= 38$, $p = 0{\cdot}24$). One of these subpopulations had an estimated median cell volume of $e^{\mu} = 93{\cdot}7$, whereas the second subpopulation had an estimated median cell volume of $e^{\mu} = 97{\cdot}5$, perhaps reflecting the appearance of reticulocytes in response to developing anaemia, as indicated by the decrease in haemoglobin detected one week earlier, since reticulocytes (young red blood cells) have a relatively increased volume in comparison with older red blood cells. As shown in Figures 1 and 2, a decrease in the mean haemoglobin, detected at the fourth week of the study, preceded the first sequential change in MCV (week 8). This phenomenon was observed in 8 of 11 patients and is consistent with our previous clinical observations. However, for this subject the change in the red blood cell volume distribution, detected at week 5, also occurred before the first significant decrease in the MCV, shown at week 8. In 6 of 11 subjects, the demonstration of multiple populations of cells in the distribution of red blood cell volume preceded detected changes in the MCV. This is a new observation resulting from this study.

4. DISCUSSION

The need for statistical surveillance of individuals in various areas of medicine has been described by Frisen.[26] Two general approaches to statistical surveillance, in situations where the number of observations is successively increasing and successive decisions are required, include the areas of quality control charts and change point analysis. General descriptions of relevant theory and methodology related to these two approaches may be reviewed in Wetherill and Brown[27] and

Zacks.[28] Some specific methods for testing a sequence of observations for a single shift in location are given by Hawkins,[7, 8] James *et al.*[9] and Hinkley and Schechtman.[10] These methods are not appropriate, however, for analysis of sequential blood test values when multiple shifts in the mean can occur. Yin[11] gives an algorithm for estimation of the number, locations, and magnitudes of jumps in a stochastic process, and Hawkins[12] describes a statistical approach to retrospective testing for parameter shifts in a linear model and explores the power for the single-shift alternative. Although these papers may have potential applications, they do not evaluate sequential application to real data with limited observations for detection of multiple shifts in location, as is the case in our study. Thompson[29] describes application of forward stepwise multiple regression techniques to population data to detect time trends in a measure of platelet aggregation. The data arose from a prospective study of the role of the haemostatic system in ischaemic heart disease. To estimate change points, they evaluate successive multiple regression models that include subsets of qualitative (dummy) variables representing a change in the mean at specified dates. Their stopping criteria was based upon the overall F-statistic. Weisberg[30] recommends criteria-based subset selection and notes that the model selected in a stepwise fashion may not optimize suitable criterion functions and may overstate significance of results. To avoid these difficulties, we used an all-possible-subsets approach by leaps and bounds with the weighted minimum risk criteria used for final model selection. We found that this method for statistical surveillance could detect sequential changes in the mean for individual laboratory tests throughout the development of iron-deficiency anaemia.

For individual blood samples, current haematology analysers are capable of performing haematological measurements such as haemoglobin, haematocrit, and mean cell volume. In addition, the distribution of the volume of red blood cells often is derived, providing the red cell volume distribution width (RDW), a calculated value that approximates the coefficient of variation of the distribution. In established iron-deficiency anaemia, characteristic changes in tests such as the MCV and RDW are helpful in suggesting the diagnosis, but the first detectable change after iron stores have become exhausted is a decrease in Hb or Hct. Decreases in serum ferritin and transferrin saturation, indicators of iron stores, are useful in the early detection of developing iron deficiency, but these tests are not performed routinely in the absence of anaemia. In the current study, we found that a statistically significant decrease in the Hb or Hct could be detected before changes in the MCV or RDW.

Statistical methods developed for analysis of the distribution of red blood cell volumes have demonstrated that in healthy individuals a log-normally distributed population of cells is present. In some patients with severe iron-deficiency anaemia, a single log-normal population of cells is also present but with decreased mean and increased standard deviation when compared to reference values from healthy individuals.[5, 19] After iron therapy, a new subpopulation of cells within the range for normals appears and the volume distributions eventually become bimodal, fitting a mixture of two log-normal distributions.[21] The importance of the current study is that it is the first to use distribution modelling to quantify sequential changes in red blood cell volume distributions during the development of iron-deficiency anaemia. We found that, for 6 of 11 subjects, alterations in the red cell volume distribution preceded alterations in the mean corpuscular volume, suggesting that this approach may provide an 'early warning' of changes in red cell production, destruction, or loss.

The development of statistical methods for deriving patient-specific reference values makes possible automated examination of laboratory data, with rapid and reliable identification of patients whose haematologic measurements have significantly changed from past values. Sequential analysis

of the red cell volume distributions may provide an early and sensitive indication of microcytic or macrocytic erythropoiesis. The methods are of general applicability to analysis of serial data from patients with other types of anaemia. By providing the statistical foundation for the automated review of laboratory data using patient-specific reference values, evaluation of test results by physicians should be facilitated through early, sensitive and reliable identification of significant changes from past values in each patient.

ACKNOWLEDGEMENTS

This work has been supported in part by an Academic Research Enhancement Award, R15 HL48349, from the National Institutes of Health, a Senior International Fellowship 1, F06 TW02117, awarded by the Fogarty International Center of the National Institutes of Health, and an Ethyl Raybould Fellowship awarded by the Department of Mathematics at the University of Queensland (CEM). Additional support was provided by funds from the Department of Veterans Affairs (GDM). We thank Emily Nielsen, R.N., of the USDA, ARS, Grand Forks Human Nutrition Research Center for recruiting volunteers and the Laboratory Service at the Fargo VA Medical Center for help with analysis of blood samples. Mention of a trademark or proprietary product does not constitute a guarantee of warranty of the product by the United States Department of Agriculture and does not imply its approval to the exclusion of other products that also may be suitable. U.S. Department of Agriculture, Agricultural Research Service, Northern Plains Area is an equal opportunity/affirmative action employer and all agency services are available without discrimination.

REFERENCES

1. Cotlove, E., Harris, E. K. and Williams, G. Z. 'Biological and analytic components of variation in long-term studies of serum constituents in normal subjects. III. Physiological and medical implications', *Clinical Chemistry*, **16**, 1028–1032 (1970).
2. Williams, G. Z., Widdowson, G. M. and Penton, J. 'Individual character of variation in time-series studies of healthy people. II. Differences in values for clinical chemical analytes in serum among demographic groups, by age and sex', *Clinical Chemistry*, **24**, 313–320 (1978).
3. Winkel, P. and Statland, B. E. 'Using the subject as his own referent in assessing day-to-day changes of laboratory test results', in Hercules, D. M., Hieftze, G. M., Snyder, L. R., *et al.* (eds), *Contemporary Topics in Analytical and Clinical Chemistry*, vol. 1, Plenum, New York, 1977, p. 287.
4. Ross, D. W., Lanier, H. A., Watson, J. and Bentley, S. A. 'Stability of haematologic parameters in healthy subjects', *American Journal of Clinical Pathology*, **90**, 262–267 (1988).
5. McLaren, C. E., Brittenham, G. M. and Hasselblad, V. 'Statistical and graphical evaluation of erythrocyte volume distributions', *American Journal of Physiology*, **252**, (*Heart Circulation Physiology* **21**), H857–H866 (1987).
6. McLaren, C. E., Houwen, B., Koepke, J., Rowan, R. M., Ortner, B. A. and Bishop, M. L. 'Analysis of red blood cell volume distributions using the ICSH reference method: detection of sequential changes in distributions determined by hydrocynamic focusing', *Clinical and Laboratory Haematology*, **15**, 173–184 (1993).
7. Hawkins, D. M. 'Testing a sequence of observations for a shift in location', *Journal of the American Statistical Association*, **72**, 180–186 (1977).
8. Hawkins, D. L. 'A simple least squares method for estimating a change in the mean', *Communications in Statistics –Simulation*, **15**, 655–679 (1986).
9. James, B., James, K. L. and Siegmund, D. 'Tests for a change point', *Biometrika*, **74 (1)**, 71–83 (1987).
10. Hinkley, D. and Schechtman, E. 'Conditional bootstrap methods in the mean-shift model', *Biometrika*, **74 (1)**, 85–93 (1987)

11. Yin, Y. Q. 'Detection of the number, locations, and magnitudes of jumps', *Communications in Statistics – Stochastic Models*, **4**, 445–455 (1988).
12. Hawkins, D. L. 'A U-I approach to retrospective testing for shifting parameters in a linear models', *Communications in Statistics – Theory and Methods*, **18**, 3117–3134 (1989).
13. Furnival, G. M. and Wilson, R. W. Jr. 'Regression by leaps and bounds', *Technometrics*, **16**, 499–511 (1974).
14. Eubanks, R. L. *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Inc., New York, 1988.
15. Gasser, T., Sroka, L. and Jennen-Steinmetz, C. 'Residual variance and residual pattern in nonlinear regression', *Biometrika*, **73**, 624–633 (1986).
16. Gallagher, S. K., Johnson, L. K. and Milne, D. B. 'Short-term and long-term variability of indices related to nutritional status. I: Ca, Cu, Fe, Mg, and Zn', *Clinical Chemistry*, **35**, 369–373 (1989).
17. International Committee for Standardization in Haematology. 'ICSH recommendations for the analysis of red cell, white cell, and platelet size distribution curves: I. General principles', *Journal of Clinical Pathology*, **35**, 1320–1322 (1982).
18. International Committee for Standardization in Haematology. 'ICSH recommendations for the analysis of red cell, white cell, and platelet size distribution curves. Methods for fitting a reference distribution and assessing goodness-of-fit', *Clinical and Laboratory Haematology*, **12**, 417–431 (1990).
19. McLaren, C. E., Brittenham, G. M. and Hasselblad, V. 'Analysis of the volume of red blood cells: application of the expectation-maximization algorithm to grouped data from the doubly-truncated lognormal distribution', *Biometrics*, **42**, 143–158 (1986).
20. McLachlan, G. J. and Basford, K. E. *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.
21. McLaren, C. E., Wagstaff, M., Brittenham, G. M. and Jacobs, A. 'Detection of two component mixtures of lognormal distributions in grouped doubly-truncated data: analysis of red blood cell volume distributions', *Biometrics*, **47**, 607–622 (1991).
22. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm (with discussion)', *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38 (1977).
23. McLachlan, G. J., McLaren, C. E. and Matthews, D. 'An algorithm for the likelihood ratio test of one versus two components in a normal mixture model fitted to grouped and truncated data', *Communications in Statistics – Simulation and Computation*, **24**, 965–995 (1995).
24. Jones, P. N. and McLachlan, G. J. 'Maximum likelihood estimation from grouped and truncated data with finite normal mixture models', *Applied Statistics*, **39**, 273–282 (1990).
25. McLachlan, G. J. 'On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture', *Applied Statistics*, **36**, 318–324 (1987).
26. Frisen, M. 'Evaluations of methods for statistical surveillance', *Statistics in Medicine*, **11**, 1489–1502 (1992).
27. Wetherill, G. B. and Brown, D. W. *Statistical Process Control*, Chapman and Hall, London, 1990.
28. Zacks, S. 'Survey of classical and Bayesian approaches to the change-point problem: fixed sample and sequential procedures of testing and estimation', in Rizvi, M.H. (ed.), *Recent Advances in Statistics*, Academic Press, New York, 1983, pp. 245–269.
29. Thompson, S. G. 'A method of analysis of laboratory data in an epidemiological study where time trends are present', *Statistics in Medicine*, **2**, 147–153 (1983).
30. Weisberg, S. *Applied Linear Regression*, Wiley, New York, 1988.

