Theory and Practice of Mixed Models Applied to Medical Research

Helen K. Brown

PhD by Research Publications
University of Edinburgh
2006

Declaration

I declare that,

a) I have composed the review section of the thesis

b) I have contributed to the research publications as indicated in the Contents list.

Abstract

This thesis examines in depth the properties of mixed models and considers their application in a variety of designs used in medical research. Mixed models are a broad class of models which allow variation in the data to be modelled at several levels and take into account correlations occurring between observations. They offer several potential advantages over the more conventional fixed effects approaches: more efficient estimates, effective handling of missing data and more appropriate inference. The different types of mixed model are placed into a unified format and the properties of various fitting methods, including likelihood-based methods, least squares methods and the Bayesian approach, are considered in detail. The practical implications of using mixed models are examined and the submitted material would appear to be the first to consider these in such depth. The particular features of applying mixed models to a range of designs are considered including repeated measures, crossover, multi-centre, meta analysis, cluster randomised, hierarchical, bioequivalence and several more ad hoc designs. Novel approaches are introduced for sample size estimation and for analysing crossover designs with multiple periods, bioequivalence studies and case-control studies. Comparisons of mixed models with fixed effects models, which have often previously been the conventional approach, are given particular attention. Models suitable for both normal and non-normal data are considered and examples involving original analyses are used to illustrate the properties described. The published material comprises two editions of a textbook and ten journal publications.

Contents

The candidate's contribution to research publications is coded below as:-

(a) primarily the work of the candidate

(b) equal contribution by the candidate and other authors, or where the candidate is responsible only for statistical analysis and interpretation.

Additionally a statement describing the candidate's contribution to the textbook (publications 1 and 1a) is provided following the publications list.

1. Brown,H.K., Prescott,R.J. *Applied mixed models in medicine*, second edition, John Wiley, Chichester, 2006. (a, except where indicated in review)

1a. Brown,H.K., Prescott,R.J. *Applied mixed models in medicine*, John Wiley, Chichester, 1999. (a, except where indicated in review)

2. Grigor,P.N., Cockram,M.S., Steele,W.B., Le Sueur,C.J., Forsyth,R.E., Guthrie,J.A., Johnson,A.K., Sandilands,V., Reid,H.W., Sinclair,C. and **Brown,H.K**. 2001. Effects of space allowance during transport and duration of mid-journey lairage period on the physiological, behavioural and immunological responses of young calves during and after transport. Animal Science, 2001, 73:341-360. (b)

3. Alexander,F.E., Anderson,T.J., **Brown,H.K.**, Forrest,A.P.M., Hepburn,W., Kirkpatrick,A.E., Muir,B.B., Prescott,R.J., Smith,A., Warner,J. 'The Edinburgh randomized trial of breast-cancer screening - results after 14 years of follow-up', 1999, Lancet (b)

4. Alexander,F.E., **Brown,H.K.**, Prescott,R.J. 'Improved classification of socio-economic status explains differences in all-cause mortality in a randomised trial of breast cancer screening, 1998, *Journal of Epidemiology and Biostatistics*, 1998, 3, 219-224 (b)

5. Luis-Fuentes,V., Moran,L., Schber,K., Dukes-McEwan,J., **Brown,H.**, Sutherland,G.R., McDicken,W.N. 'Measurement of cyclic variation in ultrasonic integrated backscatter in continuously unsedated clinically normal dogs', *American Journal of Veterinary Research*, 1997, 58, 1055-59 (b)

6. Brooke,H., Gibson,A., Tappin,D., **Brown,H.** 'Case-control study of sudden infant death syndrome in Scotland, 1992-1995', *British Medical Journal*, 1997, Vol.314, 1516-1520 (b)

7. Dixon,J.M., Ravisekar,O., Cunningham,M., Anderson,E.D.C, Anderson,T.J., **Brown,H.K.** 'Factors affecting outcome of patients with impalpable breast cancers detected by breast screening', *British Journal of Surgery*, 1996, 83, 997-1001 (b)

8. Dallak,M., Luan,X.J., **Brown,H.,** Pirie,L. 'Stability of breathing patterns in men, rats and rabbits' *Journal of Physiology - London*, 1995, 483, 96-97 (b)

9. Humphreys,F., Symons,J.A., **Brown,H.K.**, Duff,G.W., Hunter,J.A.A. 'The effects of gamolenic acid on adult atopic eczema and premenstrual exacerbation of eczema', *European Journal of Dermatology*, 1994, 4, 598-603, (b)

10. Brown,H.K., Kempton,R.A., 'The application of REML in clinical trials', *Statistics in Medicine*, 1994, 13, 1601-1617, (a)

11. Alexander,F.E., Anderson,T.J., **Brown,H.K.**, Forrest,A.P.M., Hepburn,W., Kirkpatrick,A.E., McDonald,C., Muir,B.B., Prescott,R.J., Shepherd,S.M., Smith,A., Warner,J. 'The Edinburgh randomised trial of breast-cancer screening - results after 10 years of follow-up', *British Journal of Cancer*, 1994, 70, 542-548 (b)

Note. Permission has been sought from publishers to include copies of the research publications within the submission and gained from all responding publishers.

Statement describing the candidate's contribution to the textbook

This statement describes the evolution of both editions of the textbook (references 1 and 1a) and clarifies the contribution made by the candidate.

The first edition of the textbook evolved from a constantly changing set of course notes that accompanied a three day course taught by the both authors entitled 'Mixed models analysis of medical data: PROC MIXED and beyond'. The course was first presented in 1993 and a comprehensive set of accompanying notes (approx 350 pages) were prepared by the candidate with editing and suggestions from Robin Prescott. The notes were revised with each rerun of the course to incorporate: general improvements, comments received from delegates on the course, new software and new examples. The revisions were made by the candidate.

In 1996 John Wiley and Sons agreed to publish a textbook to be developed from the course notes. The first draft was produced by the candidate with the exception of Chapter 7 and Section 6.6.1. Chapter 7 was drafted based on the material in the crossover chapter of the course notes and Section 6.6.1 contained an example that was first introduced in the textbook. Several edits of the draft were undertaken by the candidate taking into account suggestions from both authors. The notes for the 1997 and 1998 courses were based on the textbook draft and comments received during the course were also taken into account. Additionally Chapter 1 was redrafted following comments from the series editor, Vic Barnett. The redraft was undertaken by the candidate with the exception of Section 1.1 (The Use of Mixed Models).

In 2004 John Wiley and Sons invited the authors to write a second edition of the textbook. This edition updated the material in the light of new advances and comments on the first edition, reworked all examples using the latest SAS software and added new sections. Both authors worked on the revision with the candidate taking the lead on Chapters 1-4 and 9 and Robin Prescott on Chapters 5-8. Several new sections were added or substantially expanded - first drafts of Section 8.15 (bioequivalence), the majority of Section 8.16 (cluster randomised studies) and Section 9.1 (software for mixed models) were written by the candidate; while first drafts of Section 8.14

(factorial design example), 8.16.1 (cluster randomised trial example) and an expansion of Section 2.4.7 (missing data) were written by Robin Prescott. There followed several edits of the draft taking into account suggestions from both authors.

1. Introduction

This chapter includes an introduction to the thesis in Section 1.1 followed by an introduction to mixed models in Section 1.2.

1.1 Introduction to thesis

This thesis examines in depth the properties of mixed models and their application to a variety of data structures. Mixed models are a broad class of models which allow variation in the data to be modelled at several levels and take into account correlations occurring between observations. They offer several potential advantages over the more conventional fixed effects approaches: more efficient estimates, more effective handling of missing data and more appropriate inference. Models for both normal and non-normal data are considered.

The thesis consists of published material contained in two editions of a textbook and ten journal publications accompanied by a review which together constitute a PhD by research publications. The remaining material in the review is broadly split into two parts. Chapter 2 considers the properties of methods for fitting mixed models and the practical implications of their application and Chapters 3-8 consider the properties of mixed models for a variety of designs used in medical research. Particular attention is paid to comparisons of mixed models with fixed effects models which have often previously been the conventional approach. Many original analyses are included to illustrate the properties described. The designs considered primarily relate to the medical field but are equally applicable in other fields of application.

The textbook forms the major part of the published material and two editions are submitted. All direct references to chapters or sections in the review correspond to the second edition of the textbook unless otherwise stated. However, the first edition is additionally included to demonstrate the originality of some of the material at the time of first publication. Due to the length of the textbook, only sections referenced in the review form part of the submission and, in particular, sections specifying SAS code and output are not included. One journal publication is statistical (publication 10) and the remaining nine are publications in medical and veterinary journals where the applicant has contributed the statistical analysis and conclusions. These publications are used to provide examples and are in some cases contrasted to re-analyses within the textbook.

The published material referenced in each chapter of the review is listed below.

Review Chapter	***Published material***
1.1 Introduction to mixed models	1 (Chapter 1)
2. Theoretical considerations	1 (Chapters 2, 3 and 4) and 10
3. Multi-centre trials	1 (Chapter 5) and 10
4. Cluster randomised trials	1 (Section 8.16), 3, 4 and 11
5. Repeated measures designs	1 (Chapter 6, and 8.13), 9 and 10
6. Crossed designs	1 (Chapter 7) and 10
7. Hierarchical designs	1 (Sections 8.7, 8.8, 8.9, 8.10, 8.11, 8.12), 5, 7 and 8
8. More complex designs	1 (Sections 8.1, 8.2, 8.3, 8.4, 8.15) and 2
9. Discussion	Various sections of 1

1.2 Introduction to mixed models

Benefits can often be gained by using a mixed model compared to the more conventional fixed effects approaches. The advantages may include more efficient estimates, effective handling of missing data and sometimes a wider generalisation. The benefits occur because mixed models take into account correlations between groups of observations. These correlations can be modelled either by fitting random effects or coefficients to allow random variation at several levels, or by specifying correlation structures for groups of observations. Fitting random effects or coefficients causes correlations to be induced between observations within the same random effect category. For example, if patients are fitted as random in a cross-over trial, correlation is induced between observations on the same patient. Alternatively in a repeated measures trial a correlation pattern across timepoints may be specified for observations on the same patient. Examples are given in Section 1.1 to illustrate these properties.

Section 1.2 introduces a mixed model from first principles by developing it from a very simple fixed effects model and describes the main differences between fixed and random effects approaches.

In Section 1.3 a trial of treatments for hypertension is introduced. The trial was a randomised double blind comparison of three treatments for hypertension and is reported in Hall et al. (1991). One treatment was a new drug (A) and the other two (B and C) were standard drugs for controlling hypertension (A = Carvedilol, B = Nifedipine, C = Atenolol). Twenty-nine centres participated in the trial and two pre-treatment and four post-treatment visits were made. Two hundred and eighty eight patients were randomised to receive one of the three treatments and thirty patients dropped out during the study. Measurements of blood pressure, laboratory values and adverse events were recorded at each visit. Diastolic and systolic blood pressure were the primary outcome variables. The trial had been previously analysed in Hall *et al.* (1991) using a fixed effects approach. In Sections 1.3 and 1.4 analyses of the trial are used to illustrate the features of the three main types of mixed model – random effects models, random coefficients models and covariance pattern models.

Section 1.5 and publication 10 (Section 1) consider the key distinguishing features of mixed models in more detail, the benefits that can be gained from them, their particular application in medicine and a history of their use.

Definitions of containment, balance and error strata are given in Section 1.6. While there are fairly standard definitions for containment and error strata, balance in mixed models is not straightforward to define. Two definitions of balance are proposed, complete balance and balance across random effects, and the different model properties resulting from each definition are considered. Most other texts have provided a simple definition of balance as arising when there are an equal number of observations per 'cell' (where a cell relates to each combination of effects fitted) or omit to provide a specific definition. The published material would appear to be novel in providing definitions of balance for the mixed model and in examining the effect of each form of balance on results.

2. Theoretical Considerations

This chapter considers the methods and properties of mixed models. Section 2.1 considers models for normally distributed data, Section 2.2 considers generalised linear mixed models (GLMMs) which are suitable for analysing many types of non-normally distributed data and Section 2.3 considers mixed models for categorical data. The textbook (publication 1) provides a comprehensive comparison of the different fitting methods available and a rigorous evaluation of the pros and cons of using mixed models in practice. It forms the first assessment made of mixed models in such detail.

2.1 Normal mixed models

In this section mixed models with normally distributed errors are considered. While these are referred to as 'normal mixed models', it is not necessarily implied that values of the response variables follow a normal distribution. The published material relates to Chapter 2 and to publication 10. It includes a specification of the normal mixed model using a general matrix notation which can be used for all types of mixed model. This notation allows an easier understanding of the overall theory underlying mixed models and the methods used to fit them. Fitting methods for mixed models based on both classical statistical techniques and Bayesian approaches are defined and contrasted. Practical issues relating to the use and interpretation of mixed models are considered, and these are illustrated in a worked example.

The normal mixed model is specified using a general matrix notation in Section 2.1 and the covariance matrix structure is defined for the three most common types of mixed model – the random effects model, the random coefficients models and the covariance pattern model. Covariance pattern models are defined as models which have correlated residuals. These models are often used to analyse repeated measures data.

Section 2.2 and publication 10 (Section 1) describe a variety of methods for fitting the normal mixed model based on likelihood and iterative generalised least squares approaches. A mixed model is defined using fixed effects, random effects and covariance patterns. Fixed effects are not assumed to vary randomly, for example treatment effects in a clinical trial would normally be assumed to be fixed, whereas random effects are assumed to be a sample from a population. For example, the patients recruited to a clinical trial could be considered a sample from the population of interest and hence fitted as random. Covariance patterns directly define the covariance structures within groups of observations, for example repeated measurements on the same patient. Likelihood functions are defined in terms of the fixed effects and variance component parameters. The residual maximum likelihood (REML) function is derived from the ordinary maximum likelihood function and the property of unbiased variance components is shown. The alternative method of iterative generalised least squares is described and shown to lead to the same results as likelihood-based approaches, provided the same assumptions are made. In particular, iterative generalised least squares leads to the same results as ordinary maximum likelihood where the estimated fixed effects are assumed to be constants rather than unknown parameters when estimating the variance components, and restricted iterative generalised least squares leads to the same results as REML where fixed effects are assumed to be unknown parameters.

Derivations of closed form expressions for the fixed effects estimates and their variances are obtained by differentiating the log likelihood. It is shown that the expressions for these effects are identical regardless of the fitting method used. Random effects are assumed to be realisations of a distribution and hence have zero estimates. However, predictions of random effects can be formed by redefining the likelihood function so that it is conditioned on the random effects and from this a closed form expression for the random effects solutions and their variances is obtained.

In Section 2.3 the Bayesian approach to fitting mixed models is considered. The section includes an introduction to the approach and to the simulation methods that are often used to fit the models. Comparisons are made to classical approaches and it is shown that the posterior distribution is analogous to the REML likelihood when flat priors are used, and furthermore that this is almost the case when other uninformative priors are used. The potential advantages of fitting a Bayesian model arise because the distributions of the model parameters are fully evaluated providing exact estimates, standard errors and confidence regions. This contrasts with classical approaches where standard errors and confidence intervals are calculated based on point estimates and are often subject to biases (see below).

In Section 2.4 the general properties and potential deficiencies of mixed models are explored. Negative variance component estimates are considered in Section 2.4.1. The probability of obtaining a negative estimate by chance is defined and calculated over a range of scenarios. Approaches to handling a negative estimate are described and it is suggested that negative estimates are usually fixed at zero or removed from the model. However, an alternative interpretation of negative correlation between observations within random effect categories is also considered. This can be allowed for by redefining the model as a covariance pattern model where negative correlation is allowed within the categories.

Section 2.4.2 considers the importance of estimating variance parameters with a reasonable accuracy and suggests that an effect should have a minimum of five categories to be fitted as random and should otherwise be fitted as fixed.

Section 2.4.3 describes the bias that occurs in fixed and random effect standard error estimates and defines situations where it is likely to be most noticeable. The amount of bias is related to the precision of the variance component estimates, the size of the variance components and the degree of imbalance in the data. It is recommended that standard errors are adjusted using the approach suggested by Kenward and Roger (1997).

Section 2.4.4 defines formulae for Wald F and t tests for testing contrasts of fixed effect estimates and random effect predictions using matrix notation. The degrees of freedom for these tests need to reflect the combination of error strata on which the test contrast is based and methods for their calculation are described. The use of Bayesian models to provide probability values equivalent to a two-sided classical test is demonstrated. Approaches for testing the significance of variance parameters are also suggested. Confidence intervals are defined in Section 2.4.5 following the standard method of calculation but using the mixed models degrees of freedom.

The assumptions made when fitting a normal mixed model are specified in Section 2.4.6. Approaches to model checking using normal and residual plots corresponding to each error strata are described. For models with correlated residuals it is recommended that the residuals are standardised using Cholesky's decomposition matrix before being checked.

Section 2.4.7 considers the implications of using mixed models when there are missing data. In many situations mixed models are able to overcome the problems caused by missing data. However, this is based on the assumption that the data are missing at random. Three recognised types of missing data are described: missing completely at random, missing at random and missing not at random. Suggestions are made of how to categorise missing data and how situations where data are not missing at random can be handled.

A worked example of data from the hypertension trial described in Section 1.3 is provided in Section 2.5. Analyses fitting a variety of alternative models are described including models using a Bayesian approach. These demonstrate some of the features defined for mixed models and also consider each of the aspects covered in Section 2.4.

2.2 Generalised linear mixed models

Generalised linear mixed models (GLMMs) are suitable for analysing non-normal data provided it can be assumed to follow a distribution from the exponential family. Random effects, random coefficients or covariance patterns can be included in a

GLMM in much the same way as in normal mixed models, and again either balanced or unbalanced data can be analysed. Chapter 3 defines the GLMM using a general matrix notation, building the model from its fixed effects counterpart the generalised linear model (GLM). Alternative methods for fitting the GLMM are defined and contrasted. Aspects relating to practical application are considered and a worked example is provided.

In Section 3.1 the fixed effects GLM is introduced. Definitions of common distributions are provided along with a specification of the general form for distributions from the exponential family. The GLM is specified using matrix notation and a fitting method based on maximum likelihood is illustrated. The conditional logistic regression approach is also defined as an alternative fixed effect approach.

Section 3.2 specifies the extension of the GLM to a GLMM using matrix notation. The likelihood function is defined for random effects models and a quasi-likelihood for situations where a true distribution cannot be defined. Fitting methods corresponding to pseudo-likelihood, generalised estimating equations and Bayesian methods are introduced and their relative benefits are discussed.

The practical application and the implications of using GLMMs are considered in Section 3.3. The material would appear to be the first to consider these aspects in such depth.

Section 3.3.1 considers the alternative ways in which binary data can be analysed. The section concludes that analyses based on Bernoulli rather than binomial data are preferable for fitting GLMMs. Analysis of the data in this form allows for a more appropriate modelling of the variance structure and also allows covariates recorded at the observation level to be fitted.

Section 3.3.2 considers the effects of 'uniform' categories. A category is defined as uniform when all observations within it have the same value, for example a patient effect category in a cross-over trial is uniform if all results for the patient are the same. While a uniform category causes infinite effect estimates to occur in a conventional GLM, often satisfactory estimates can be obtained when the effect is modelled as random in a GLMM. However, if too many categories are uniform fitting problems can also arise in the GLMM.

Section 3.3.3 considers negative variance component estimates. The same considerations apply as in the normal mixed model, however potential biases caused by the fitting methods can make a negative variance component estimate even more likely.

Section 3.3.4 describes how fixed and random effect estimates can be meaningfully expressed by applying the inverse of the link function to the estimates obtained on the linear scale from the GLMM.

The accuracy of variance parameters is considered in Section 3.3.5. As for normal data it is suggested that a reasonable accuracy is required because the variance components are used to estimate both effect means and their standard errors. At least five categories for an effect are again recommended to fit it as random. The potential bias of variance parameters in GLMMs caused by random effects shrinkage, and in particular in the presence of uniform categories, is described.

Section 3.3.6 considers biases in fixed and random effects standard errors. These are biased both for the same reason as in normal mixed models (see Section 2.1 of review), and also due to biases in the variance component estimates. Biases for the first reason can again be overcome by using the adjustment suggested by Kenward and Roger (1997).

The effects of including a dispersion parameter in the model are considered in Section 3.3.7. Inclusion of the parameter helps to overcome biases in the variance component estimates and in some situations allows over-dispersion in the data to be modelled. It is concluded that the parameter should always be included in GLMM analyses.

Significance tests for GLMMs are considered in Section 3.3.8 and Wald F and t tests based on the linear effects estimated from the GLMM are recommended. Confidence intervals are defined in Section 3.3.9 and follow the conventional approach.

An illustrative example of a binary variable from the hypertension trial (see Section 1.2 of review) is provided in Section 3.4. Analyses fitting a variety of alternative models are carried out including models using the Bayesian approach. These demonstrate some of the features defined for GLMMs and also consider each of the methodological aspects described above and covered in Section 3.3.

2.3 Mixed models for categorical data

Categorical data often occur in clinical trials, for example adverse events may be classified on an ordinal scale as mild, moderate or severe. The published material describes the fixed effects ordinal logistic regression model and extends this to mixed models for analysing ordinal and unordered categorical data. Fitting methods and the practical implications of fitting the models are considered, followed by a worked example.

The ordinal logistic regression model is defined in Section 4.1. An alternative specification is also provided in matrix notation using an extended binary format for the multinomial observations. This allows the correlations occurring between the multinomial categories to be defined in a covariance matrix.

Section 4.2 specifies the ordinal mixed model using the extended binary format following the specification given by Lipsitz et al. (1994). Structures for covariance matrices are provided taking account of random effects, covariance patterns and the multinomial correlations. Likelihood and quasi-likelihood functions are defined. The methods available for fitting categorical mixed models are summarised.

A mixed model for unordered categorical data is specified in Section 4.3. This has a similar form to the mixed model for ordinal data except that parameters corresponding to each category are now needed to model each fixed and random effect. For variance components a separate parameter is needed for each pair of categories.

The practical implications of fitting mixed models to categorical data are considered in Section 4.4. The proportional odds assumption made in analyses of ordinal data is described in Section 4.4.1. This assumption is made because the same fixed effect parameter is used to model all partitions of the categories. It is concluded that the assumption is usually appropriate as different fixed effects for each partition would be difficult to interpret and usually an overall estimate across partitions is of most interest. However, the partitions can be modelled separately in situations where each partition of categories may give rise to a different interpretation.

In Section 4.4.2 the requirement for a larger number of covariance parameters required compared to an equivalent GLMM is discussed. It is recommended that more complex terms in the covariance matrix are introduced cautiously. Section 4.4.3 recommends that a covariance pattern is chosen by carrying out likelihood ratio tests to test whether more complex patterns can be justified. The difficulty in interpreting multiple covariance parameters is discussed in Section 4.4.4.

A worked example is provided in Section 4.5. This is based on an analysis of an adverse event 'cold feet' recorded in the hypertension study (see Section 1.1 of review). Cold feet were recorded on an ordinal scale of 1– 5: 1 = none, 2 = occasionally, 3 = on most days, 4 = most of the time, 5 = all of the time. The analyses demonstrate the use of alternative covariance structures to model the repeated measurements.

3. Multi-Centre Trials and Meta-Analyses

The analysis of data that are collected from several centres or trials is considered in this chapter. A multi-centre trial is carried out using several centres either because insufficient patients are available for any one centre, or with the deliberate intention of assessing the effectiveness of treatments in several settings. The published material relates to Chapter 5 and publication 10. The statistical and interpretational implications of fitting mixed models to multi-centre data compared to fixed effects approaches are considered in depth and the alternative inferences that can be made from a range of models are described. Two worked examples are included. Sample size formulae are derived corresponding to the mixed model fitting centre and centre.treatment effects as random. This would appear to be the first time such formulae have been specified. Meta-analysis data have the same basic structure as data from multi-centre trials and it is demonstrated that similar mixed models can be applied.

Sections 5.1 and publication 10 (Section 4) provide an introduction to multi-centre trials and their analysis using mixed models. Sometimes there will be extra variability in a multi-centre trial due to differences between the centres (e.g. different investigators, types of patients, climates). This extra variation can be taken into account in the analysis by including centre and centre.treatment effects in the model. In most situations a more efficient analysis will be obtained by fitting centre effects as random. However, deciding whether centre.treatment effects should be fixed or random can be the subject of debate. In practice, the choice will depend on whether treatment estimates are to relate only to the set of centres used in the study or, more widely, to the circumstances and locations of which the trial centres can be regarded as a sample. In the former case, ‘local’ treatment estimates for the sampled set of individual centres are obtained by fitting centre.treatment effects as fixed. To obtain ‘global’ treatment estimates with a wider inference, centre.treatment effects should be fitted as random. When this is done the standard error of treatment differences is increased to reflect the heterogeneity of the treatment effects across centres.

The potential reasons for extra variability in treatment effects across centres are considered. Variation in a treatment effect across centres is likely to be most noticeable in trials that do not compare drugs. For example, in a trial to compare surgical procedures there may be varying levels of experience available at each centre with the different procedures.

The statistical implications of fitting the following models are defined in Section 5.2:-

- Centre and centre.treatment effects fixed
- Centre effects fixed, centre.treatment effects omitted
- Centre and centre.treatment effects random
- Centre effects random, centre.treatment effects omitted

For each model: formulae for the variability of the treatment estimate are specified, the implications for treatment effect estimates at each centre are described, and the inferences that can be drawn are described. The section clarifies the two main purposes of using a mixed model in multi-centre trials. Firstly fitting centre effects as random provides more efficient treatment effects whenever there is imbalance in the data as occurs in most multi-centre trials. Secondly fitting centre.treatment effects as random allows a wider interpretation based on the assumption that treatment effects vary between the centres. The effect of fitting alternative models is illustrated in a worked example in Section 5.3 and in publication 10 (Section 4.1).

Section 5.4 considers in more depth the interpretation of results from alternative models and the practical aspects relating to fitting mixed models to multi-centre data. The generalisability of results is discussed depending on whether or not random centre.treatment effects are included in the model. It is concluded that generalisation often cannot be clearly defined and needs to be by degree. Several aspects need to be considered when generalising results: whether the study is single or multi-centre; the assumptions made by the model fitted; the particular application. Multi-centre studies may be considered more generalisable than studies carried out at a single centre, and results from analyses of multi-centre studies fitting centre.treatment effects as random may be considered still more generalisable than analyses that do not. An interaction between treatment and centres is more plausible in situations where variation in treatment effects across centres, for example in a trial of a surgical technique where there may be differences in surgeon experience. However, in a situation where it is less plausible for a treatment effect to vary across centres, for example in drug trials, results from analyses omitting random centre.treatment effects may still have a high level of generalisability.

Section 5.4 also considers how results are affected by the number of centres and the sizes of centres. It is suggested that as variance component estimates are used to estimate both the fixed effects and their standard errors, they should be estimated with a reasonable accuracy based on a minimum of about five centres. Mixed models are recommended even when some centre sizes are very small, perhaps containing only one patient. This avoids the approach sometimes used in fixed effects analyses where data from small centres are combined, often using an arbitrary rule to define when a centre is 'small'.

In Section 5.5 formulae for estimating sample size are derived to provide adequate power to carry out an analysis where centre.treatments are fitted as random. Formulae are also derived for non-normal data. The three situations listed below are considered:-

- Number of patients required per centre when the number of centres is specified
- Number of centres required per centre when the average centre size is known
- Number of centres and patients per centre when the relative costs of managing a centre compared to that of recruiting a patient can be estimated.

Section 5.6 demonstrates that mixed models can be applied in meta analyses with the same implications as in multi-centre trials, by replacing centre effects with trial effects. In a meta analysis a trial.treatment interaction is often more plausible than a centre.treatment interaction due to the likely differences in study protocols. The wider inference provided by a mixed model then becomes more attractive. A worked example in Section 5.7 illustrates the effects of taking alternative modelling approaches.

4. Cluster Randomised Trials

This chapter considers the use of mixed models to analyse cluster randomised trials. Cluster randomised trials are carried out when it may not be practical or ethical for subjects to receive different treatments within the same cluster (eg centre, hospital, clinic, GP practice). In this situation clusters rather than subjects are randomised to treatment groups to form a so-called 'cluster randomised' design. The published work considers the implications of mixed model approaches compared to alternative more conventional approaches. Additionally methods to classify socio-economic status to adjust for biases between groups in cluster randomised trials are considered. Data from the Edinburgh Breast Screening Trial (publications 3, 4 and 11) are used to demonstrate differences in approaches and the methods for socio-economic adjustment.

Section 8.16 describes the use of a mixed model with treatment effects fitted as fixed and cluster effects as random. The variance of the difference between treatments is given by,

$$\mathrm{var}(t_i - t_j) = \sigma^2 (1/n_i + 1/n_j) + \sigma_c^2/(1/c_i + 1/c_j)$$

where σ^2 is the residual variance, n_i and n_j are the numbers of subjects receiving treatments i and j, σ_c^2 is the centre variance component, and c_i and c_j are the number of centres allocated to groups i and j. This compares to an analysis taking no account of cluster effects where the variance would be $\mathrm{var}(t_i - t_j) = \sigma^2 (1/n_i + 1/n_j)$. It is concluded that a mixed model with cluster effects fitted as random should be the norm for analysing this type of data since treatments are estimated at the cluster level of variation. Unlike the multi-centre trial a choice of models offering alternative inferences is not available since randomisation has taken place at the cluster level. For non-normal data a GLMM fitting cluster effects as random is compared to the alternative approach of fitting a generalised linear models (GLM) with an over-dispersion parameter. The GLMM is considered preferable because it provides a more appropriate modelling of variance and also allows covariates recorded on individual observations to be fitted to help adjust for differences between groups.

In Section 8.16.2 and publications 3, 4 and 11 analyses of the Edinburgh randomised trial are considered. This was a large cluster randomised trial involving 54654 women aged between 45-64 from 87 general practices in Edinburgh. It was started in 1978 at a time when there was no breast screening programme in Scotland. Practices were randomly assigned to either an intervention or control group. Women in the intervention group were invited to participate in a screening programme involving an annual screen for breast cancer, while those in the control group received normal medical care. All subjects in the intervention practices were considered part of the intervention group, whether or not they attended for screening. The primary endpoints were death from breast cancer and death from all causes. It became apparent early in the study that there was an imbalance between the treatment groups in terms of socio-economic status and analyses have sought to adjust for this (publication 4).

Analyses of the study were carried out after 7 years (Roberts et al., 1990), 10 years (publication 3) and 14 years (publication 11) of follow up. For consistency a standard statistical approach was used across all the analyses. This involved fitting GLMs with Williams' modification (Williams, 1981) to model the extra variability (or 'over-dispersion') occurring between practices and differences between the practices were adjusted for fitting categories to represent the overall socio-economic status of each practice.

Data from the same trial were re-analysed in Section 8.16.2 using software that was unavailable at the time of the original analysis. The analyses demonstrated the benefits of fitting cluster effects as random in a GLMM rather than the alternative approach of fitting an over-dispersion parameter in a GLM. Analysing individual data in Bernoulli form was shown to be preferable to analysing summary data for each cluster in binomial form. This benefit was particularly noticeable in analyses adjusting for age and socio-economic status. This indicates that the use of summary covariate statistics for clusters can be inadequate and that fitting individual covariate values with the data analysed in Bernoulli form is likely to be preferable.

While the analyses carried out in Section 8.16.2 demonstrated that adjusting for differences in clusters using summary characteristics may be inadequate, they did not address a potential bias between the groups in the classification of socio-economic status. The bias was specific to this study. It occurred because the status was defined from each woman's postcode and postcodes of women who moved house during the study were only updated for women in the study group and not those in the control group. This bias is considered in publication 4 and a method to determine a potentially unbiased classification is specified. This new classification was calculated at the GP practice level and was used as an adjustment in the 14 year analysis (publication 11). While the approach potentially overcame the problem of bias between the groups caused by individual women's classifications, the adjustment may still have been inadequate since it was based on summary statistics and not individual classifications of socio-economic status for individual women.

5. Repeated Measures Designs

Mixed model approaches to analysing repeated measures data are considered in this chapter. Any dataset in which subjects are measured repeatedly over time can be described as repeated measures data. Repeated measurements are either made at pre-determined intervals (e.g. at fortnightly visits), or in a less controlled fashion so there are variable intervals between the repeated measurements. The type of analysis model chosen will depend on whether or not the intervals are fixed.

The published material relates to Chapter 6 and publications 9 and 10. The properties of two mixed model approaches for analysing repeated measures data are defined. These are based on either specifying a covariance structure for the repeated measurements or using random coefficients to model the repeated measurements as a function of time. Several worked examples are used to illustrate the models suggested for both normal and non-normal data. Sample size formulae are derived for the situation where it is planned to analyse the data using a covariance pattern model and the average treatment effect across visits will be of interest.

5.1 Covariance pattern models

In Section 6.1 the repeated measures design is introduced and a variety of fixed effects analysis approaches suitable for the situation where repeated measurements are made at fixed time points are described. These include the analysis of the mean response over time, separate analyses at each time point, analyses of response features and analyses of the raw data. The statistical and interpretational implications of using each of these models and the potential benefits of mixed models are described.

Covariance pattern models are defined in Section 6.2 and publication 10 (Section 3). A variety of approaches to defining the covariance structure for the repeated measurements are introduced and methods for choosing between alternative models are considered. The effects of including different fixed effects in the model are also described. In particular the inclusion of a treatment.time interaction causes an overall treatment effect estimate to be an unweighted average across time points, whereas omitting this term will result in a treatment estimate based on using different weights at each time point.

Four worked examples illustrate the properties of covariance pattern models applied to normal data, binary data, count data and ordinal data.

The first is described in Section 6.3 and is based on analysing diastolic blood pressure from the hypertension dataset measured at four post treatment visits (see Section 1.2 of review). This example considers models with a variety of covariance patterns. It illustrates the use of likelihood ratio tests to choose between models, and the effect of including and excluding the treatment.time fixed effect. A check of model assumptions is demonstrated involving use of the Cholesky decomposition matrix to take account of the correlated residuals. A complex covariance structure is selected involving separate sets of covariance parameters for each treatment group. However, estimates of treatment differences are shown to be very similar to those obtained from a simple structure with a constant correlation across treatments. It is concluded that the similarity is likely to have occurred because the amount of missing data is small and hence there is little imbalance in the data. In other less balanced datasets a greater difference in the results might have been expected.

The second example is described in Section 6.4. Data from a placebo controlled trial of an anti-convulsant treatment for epilepsy reported in Thall and Vail (1990) are analysed. The study involved fifty nine patients and the number of epileptic seizures was counted over an eight-week period prior to treatment and then over four two-week periods following treatment. No patients dropped out of the study. A histogram of the number of epileptic episodes suggested that a Poisson distribution could be used to model the data and an initial analysis was carried out using a GLMM. However, subsequent checks indicated that model assumptions were not met and the data were re-analysed using a mixed model for ordinal data after classifying the numbers of seizures into four categories.

The third example considers data from a placebo-controlled trial to assess the efficacy of gamolenic acid for treating eczema and is described in publications 9 and 10. Fifty two patients were recruited and several dropped out before the study end. Measurements were made at baseline and at six post-treatment visits. The severity of eczema is measured by a redness score covering twenty parts of the body. Fixed effects models are used to analyse data recorded at each time point in publication 9. In publication 10 analyses using covariance pattern models were carried out using the same data and these are contrasted to the fixed effects approaches previously used.

The fourth example is described in Section 8.13 and considers the situation where an effect of interest varies over time. Data were again taken from the eczema trial described above. Daily measurements of eczema were recorded in patients' diaries and women also recorded details of their menstrual cycle. These data were analysed using a covariance pattern model to assess the effect of the menstrual cycle on eczema severity and whether this interacted with treatment. The study showed subjects were more likely to experience severe itchiness during menstrual bleeding and that this was not influenced by treatment.

5.2 Random coefficients models

Random coefficients models are considered in Section 6.5. These models assume random variation in the time slopes (or curves) for each patient and provide an appropriate estimate and standard error of the differences in slopes between treatment groups. The invariance of the results to the specification of time origin is demonstrated. Approaches to handling negative variance components corresponding to random coefficients are considered and it is suggested that random coefficients are removed in order of decreasing order of complexity until all variance components are positive. Alternative approaches to using baseline measurements are considered. These can either be fitted as a covariate or included as an initial time point depending on the interpretation required.

An example comparing two treatments for HIV positive patients is described in Section 6.6.1 (b, not primarily work of candidate). The treatments are compared in terms of changes in patients' CD4 counts using a linear random coefficients model. A second example involving a polynomial random coefficients model is described in Section 6.6.2. Here measurements of herpes virus levels made on children suffering from either leukaemia or solid lump tumours were analysed. Their hospital visits did not occur at fixed intervals making the use of a random coefficients model appropriate. The models compared the differences in the pattern of change in virus levels between the two cancer groups and the data were used to demonstrate the invariance of the results to the time origin.

5.3 Sample size estimation

Sample size formulae corresponding to a covariance pattern fitting a compound symmetry structure are derived for normal and non-normal data in Section 6.7. These formulae produce smaller sample sizes than the standard parallel group calculation. While the exact covariance structure for the repeated measurements is not known in advance it is likely that the assumption of a compound symmetry structure will provide an adequate sample size estimate.

6. Crossed designs

A crossed design can be defined when the effect of interest (eg treatment) is crossed with another 'nuisance' effect (eg patients). For example, in a cross-over design each patient receives several treatments and treatment effects are then crossed with patients. Another example of this design is the matched case-control study where 'matched sets' of patients contain at least one case and control. Crossed designs have also been frequently used in crop trials in agriculture where treatments are allocated within rows of a field. The ANOVA technique and early applications of mixed models were initially developed for analysing of such experiments (for example see Talbot, 1984).

The published material concentrates primarily on cross-over trials and considers a variety of designs. The benefits of fitting patient effects as random rather than fixed are assessed for each design. For some designs this has involved the derivation of algebraic formula. A novel approach involving the use of covariance patterns to structure the correlations occurring between repeated measures by either period or treatment is also described. The matched case-control study is considered and the benefits of fitting matched set effects as random rather than fixed are shown. Practical aspects specific to crossed designs are considered and worked examples involving normal and non-normal data are provided.

6.1 Cross-over trials

The cross-over trial is introduced in Section 7.1 and publication 10. In cross-over trials, subjects are randomised to receive different sequences of treatments, with the outcome being assessed for each treatment period. Most cross-over trials have the same basic design where every subject receives each of the treatments being evaluated, for a standard period of time, with the outcome variables being assessed in the same way in each period of treatment. Examples of more complex designs include 'incomplete block' designs, where patients are randomised only to receive some of the treatments being studied, and 'optimal' designs where patients may receive the same treatment more than once.

The potential advantages of using a mixed model to analyse cross-over data are described in Section 7.2. These include more efficient estimates of treatment effects whenever there are missing data and the potential to model the covariance structure of the data when there are more than two periods in the trial.

The AB/BA cross-over trial is introduced in Section 7.3. Formulae for within- and between- patient estimates of treatment variation are specified in publication 10 and re-presented in Section 7.3. From these the relative efficiency of the mixed model estimate compared to the fixed effects model estimate is given under different circumstances. Gains in efficiency are shown to be greatest when the patient variance component is small compared to the residual, and when a larger proportion of data is missing. In order to assess the amount of gain likely in practice, the ratio of the patient variance component relative to the residual is calculated for a range of examples. A worked example is provided in Section 7.3.1 illustrating the gain in efficiency.

In Section 7.4 higher order complete block designs are considered. In these designs there are as many treatment periods as there are treatments to be compared, and each patient receives every treatment. If there are no missing data, then a conventional least squares analysis fitting treatment, period and patient effects is fully efficient. However, when there are missing data, some of the within-patient treatment comparisons are unavailable for every patient and additional between-patient information can then be utilised. The analysis of a four-period, four-treatment trial taken from Jones and Kenward (1989) is considered in Section 7.4.2. The effects of the treatments on cardiac output are compared in terms of left ventricular ejection time. Fourteen patients participated in the trial and the original data were complete. In order to demonstrate the benefits of using a mixed model thirteen observations chosen arbitrarily were set to missing. The results showed small gains in efficiency in all the mixed models. These were slightly more noticeable when carryover effects were included.

Incomplete block designs are introduced in Section 7.5. In this design the number of periods is less than the number of treatments to be evaluated so complete balance is not achieved even when data are complete. In this situation a mixed model will always provide a more efficient analysis because it utilises between-patient information in estimating treatment effects. Formulae for the variances of treatment effect estimates are derived for the three-treatment two-period design in publication 10 (and reproduced in Section 7.5.1) for models fitting patient effects as fixed or random, and for models including and excluding carryover effects. The gains in efficiency are shown to be particularly great when carryover effects are fitted. The large gains in efficiency are illustrated in a worked example using data from Mead (1988) in publication 10 and in Section 7.5.1.

Section 7.6 considers optimal designs such as Balaam's design involving the randomisation of patients to treatment sequences AA, AB, BA or BB. These designs are suitable in situations where carryover effects are expected. When carryover effects are to be included in the analysis model, this design will provide uniformly most powerful unbiased estimates of treatment and carry-over effects. Analysis of this design using a mixed model further increases the efficiency of the estimates. A trial using Balaam's design taken from Hunter et al. (1970) is analysed in Section 7.6.1 and demonstrates the benefits of using a mixed model.

The use of covariance pattern models in cross-over trials containing more than two periods is considered in Section 7.7. Models structuring the covariance by period effects and alternatively by treatment effects are specified in Sections 7.7.2 and 7.7.3. Structuring the covariance matrix by treatment allows observations on different treatments to have different variances and covariances. The models are illustrated in a worked example in Section 7.7.3 again using data from the four-period four-treatment trial considered above. Comparisons showed that a model structuring the covariance by treatment led to a better model fit. This would appear to be the first published specification and application of this approach. Since the publication of the first edition of the textbook, this approach has been more widely applied and the calculation of population bioequivalence is based on structuring the covariance by treatment (as considered in Section 8.5 of the review).

The analysis of binary data is considered in Section 7.8 and a worked example of a two-way cross-over design is provided. The benefits of the mixed model are shown compared to fixed effects approaches suggested by Prescott (1981) and Senn (1993). To avoid the potential problems caused by a large number of uniform patient effects, it is recommended that a GLMM with a compound symmetry structure is preferable to a GLMM with patient effects fitted as random. In Section 7.9 an example using a mixed model for analysing ordinal data from a two-way cross-over trial taken from Jones and Kenward (1989) is considered.

Knowledge of variance component sizes can be used in trial design to help decide whether a cross-over or between patient trial is most appropriate. This is illustrated in Section 7.10. The variance components obtained in the analysis of the oral mouthwash trial in Section 7.3.1 are used to determine the design of a future trial of mouthwashes.

General points relating to the practical application of mixed models in cross-over trials are made in Section 7.11. It is suggested that in situations where the amount of missing data is minimal, the small gain in efficiency obtained from using a mixed model may be outweighed by the benefits of using a simpler fixed effects model where there are no issues such as biased standard errors. Caution is particularly suggested when modeling non-normal data. The pros and cons of fitting carryover effects are considered and following Senn (2002) it is recommended that cross-over designs are only used when carryover effects are not expected, since these effects cannot be adequately accounted for in a statistical model.

6.2 Matched case-control studies

Matched case-control studies are considered in Section 8.5 and publication 6. In this design a group of subjects who have a particular disease or outcome (cases) are compared with a group of subjects who do not have the disease or outcome (controls). Each case is matched to one or more controls using one or more factors that are known to be connected with the disease. For example, age and sex are often used. The primary objective in a case-control analysis is to determine which factors (not used in the matching) differ between the case and control groups. However, in doing so it is important to allow for the matched nature of the data. The published material describes the analogy between the case-control design and cross-over designs, demonstrates the advantages of using a mixed model compared to the more conventional fixed effects approaches, and provides examples of analyses of normal and non-normal data.

The design of the matched case-control study is similar to that of the cross-over trial. In the cross-over trial the treatment effects are 'crossed' with patient effects (i.e. each patient may receive several treatments). In the matched case-control study, group effects (i.e. whether case or control) are crossed with matched set effects (sets of matched subjects are referred to as 'matched sets'). The effect of fitting matched sets as random in a case-control study is similar to that of fitting patients as random in a cross-over study. Results will be identical to an analysis fitting matched sets as fixed whenever there are the same number of controls for every case and the matched set variance component is positive (when it is negative and set to zero, the mean group differences will be identical but their standard errors will differ). In a fixed effects analysis fitting matched sets as fixed, information is completely lost on group effects in any matched sets which either contain only a case or only controls. Additionally, in a fixed effects analysis of binary data (which can be performed using conditional logistic regression), matched sets whose members all have identical outcomes (i.e. are uniform) do not contribute information to the analysis. This loss of information does not occur when matched sets are fitted as random because information is then 'recovered' from the matched sets error stratum. When data are non-normal the problems caused by uniform categories (see Section 2.2 of review) may occur when matched set effects are fitted as random in a GLMM. It is suggested that a GLMM with a compound symmetry covariance structure is used to avoid these difficulties.

Analyses of a study carried out by the Scottish Cot Death Trust are used to illustrate the use of mixed models and their advantages. The study aimed to interview the parents of every sudden infant death syndrome (SIDS) baby in Scotland during 1992-1995. The parents of two matched control babies born immediately before and after each case at the same hospital were also interviewed. However, not all parents agreed to participate and this caused some of the matched sets to be incomplete.

An analysis of all variables recorded in the study using conditional logistic regression analyses is reported in publication 6. Conditional logistic regression is a fixed effects approach where results are conditioned on matched set effects and is equivalent to fitting matched set as fixed. While it is a less efficient technique than using a GLMM when there are missing data, at the time of publication it was the standard approach for analysing this type of data and requested by the journal reviewers! In Section 8.5 mixed model analyses of selected quantitative and binary variables are compared to fixed effects approaches including conditional logistic regression.

7. Hierarchical Designs

Designs where data can be grouped into hierarchies and where the primary objective is not to estimate a fixed effect can be described as 'hierachical'. Often interest centres on one or more of the following:-

- estimating variance components to determine sources of variation in the data
- estimating shrunken random effects
- providing appropriate estimates of an overall mean value and standard error.

Six examples are considered. Each demonstrates the benefits of using mixed models for use with a particular data structure and compares these models to corresponding fixed effects.

Section 8.7 and publication 8 describe an experiment to calculate variance components for breathing measurements in rabbits. The inspiration times for one hundred breaths were measured on four rabbits on each of four days. This gave rise to four potential sources of variation - rabbits, days, rabbit.day interaction and residual (between breaths). A random effects model was fitted to inspiration time with each of the above effects taken as random. A positive variance component was obtained corresponding to each random effect allowing the relative sizes of variation from the different sources to be interpreted. In Section 8.7.1 the variance components are used to design a future clinical trial to compare two treatments. Sample sizes for trials with between- and within- rabbit designs are considered. Formulae are specified for the variation of the mean treatment difference in inspiration time. From this, formulae are derived for the treatment difference that can be detected for a given sample size and power, calculated over a range of sample sizes.

Data from an informal study carried out by Edinburgh radiographers to measure inter- and intra- observer variation in foetal age predicted from ultrasound scan measurements is analysed in Section 8.8. Six radiographers participated in the experiment. Fifty-two women in the latter stages of pregnancy with a mean gestation of 29.9 (SD 3.2) weeks were each scanned by two of the radiographers selected at random. Both scans were carried out in the same session. A random effects model was fitted to the data with radiographer effects taken as random and subject (women) effects as fixed. Subjects were fitted as fixed because they each have a different gestation and therefore cannot be treated as a randomly distributed sample. The radiographer variance component was zero demonstrating that there was no systematic bias in the assessments of individual radiographers. A measure of intra-observer variation was provided by the residual estimate and was used to calculate the accuracy of foetal gestation predictions.

An experiment to measure components of variation and to provide a mean estimate in a cardiology experiment on dogs is described in Section 8.9 and publication 5. The heart wall thickness of eleven healthy dogs was measured using ultrasound scans. Each scan consisted of twenty thickness measurements taken over a single heartbeat cycle. Each dog had between two and six scans and each scan was assessed by between one and three observers. Thus the data are not balanced. A random effects model with dog, observer, dog.observer and scan effects fitted as random is used to estimate variance components and to calculate an appropriate estimate for the mean and standard error of heart wall thickness. The variance components estimates allowed the sizes of variation from different sources to be assessed. The overall mean estimate weighted the observations appropriately according to the covariance structure of the data. It was shown to be quite different from a crude mean and standard error based on treating the observations as independent.

Cluster sample surveys are considered in Section 8.10. In such surveys data occur within clusters and clusters are usually sampled at random. The purpose of a cluster sample survey is to provide an appropriate estimate and standard error of a measurement. This contrasts with cluster randomised trials which seek to compare treatment effects and randomise treatments to clusters. A mixed model fitting cluster effects as random provides results that can be related with some confidence to the population of clusters. The model also produces shrunken cluster estimates. These avoid the potential problem of unrealistic estimates occurring due to chance variation when cluster sizes are small. Such an analysis using cluster sample survey data taken from Thrusfield (1995) is described in Section 8.10.1. A GLMM is applied to estimate the prevalence of a disease in animals.

A GLMM is used in Section 8.11 to provide shrunken estimates of mortality estimates for each postcode area in Edinburgh. The shrunken estimates avoid the extreme estimates that may arise in small areas.

In publication 7 data recording the outcome from mastectomy operations at an Edinburgh hospital were analysed using fixed effects GLM models. The analyses tested the effect of surgeon experience and other factors likely to affect outcome. However, the fixed effects approach was unable to provide satisfactory estimates of individual surgeon performance, particularly for a surgeon who only carried out three operations. In Section 8.12 the same data were analysed using a GLMM to provide shrunken estimates of complication rates of surgeons carrying out mastectomy operations. These shrunken estimates helped to avoid the extreme estimates that may arise for surgeons only carrying out a few operations.

8. More complex designs

The designs considered in this chapter use combinations of the features that have already been considered in the critical review – hierarchical structures, crossed designs and repeated measurements designs. Hierarchical repeated measures structures, multi-centre repeated measures data, multi-centre cross-over data, multi-centre trials with centres that are grouped hierarchically and bioequivalence studies with replicate cross-over designs, are included. The flexibility of mixed models to incorporate several design aspects into a single model is demonstrated and the practicalities connected with fitting such more complex models are considered. Comparisons are made with other more conventional methods of analysis.

8.1 Designs with repeated measurements within visits

Designs with repeated measurements occurring within visits are considered in Section 8.1. These designs can arise in both cross-over and repeated measures trials. For example, bioequivalence trials often record several blood or urine measurements at each visit in a cross-over design. Studies in cardiology sometimes involve exercise tests where repeated measurements are made throughout the test at each visit. The published material defines a variety of alternative mixed models based on covariance pattern models and random coefficients models. The properties of the alternative models are compared. Methods for model building and model selection are considered. Three worked examples illustrate the properties of the models suggested.

The use of mixed models with covariance patterns structured across visits and across repeated measurements within visit is considered in Section 8.1.1. These models are contrasted to an alternative simpler approach based on analysing summary statistics for each visit (e.g. area under the curve, maximum value or time to maximum value) with methods appropriate for ordinary repeated measures or cross-over data. While this latter approach has the advantage of simplicity and gives a straightforward interpretation, it cannot test the interaction between treatments and repeated measurements within visits or satisfactorily overcome problems caused by missing data. The mixed model takes account of missing data and allows the interaction between treatment and the 'within visit' repeated measurements to be tested. Several alternative approaches to structuring the covariance matrix are described. The two hierarchies of repeated measurements leads to a wide choice of covariance structures and care is required to ensure that parameters are not confounded. It is recommended that a model is built up from a simple structure and that more complex features are added one by one and using likelihood ratio tests to test whether they make a significant contribution.

A worked example is provided in Section 8.1.2 illustrating the use of models with covariance patterns structured both by visits and repeated measurements within visits. Data were taken from a three-period, cross-over trial taken from Jones and Kenward (1989) to compare the effects of three treatments on systolic blood pressure. There were twelve patients and ten measurements were made at each visit. The trial was previously analysed in Jones and Kenward (1989) using approaches based on ANOVA. While this is satisfactory because the data are complete, more care is required in determining the appropriate sums of squares and degrees of freedom for significance tests and for calculating the standard errors for each effect, than when a mixed model is used. This is demonstrated in the analysis carried out in Section 8.1.2. If there had been missing data in the study then an ANOVA approach would not have been appropriate.

A second example involved an experiment where repeated measurements were made within visits and is described in publication 2. This study was carried out in calves to examine the effects of space allowance during transportation. Ninety six calves were randomised to one of four groups – smaller space allowance with food and water, larger space allowance with food and water, smaller space allowance without food and water, and larger space allowance without food and water. The study involved four periods – journey 1, lairage, journey 2 and post treatment. Measurements were made at two timepoints within each period, at one and twelve hours. Mixed models were fitted with space allowance, food and water, timepoint and their interactions taken as fixed effects. In this example constant covariances across periods and across timepoints were used to provide a straightforward interpretation as many variables were analysed.

In Section 8.1.3 the use of random coefficients models to analyse data with repeated measurements made within visits is considered. Models fitting slopes across either visits, repeated measurements within visits or across both are specified. This type of model is appropriate if greatest interest is centred on explaining the relationship of a measurement with time or if either set of repeated measurements occur at irregular intervals. In Section 8.1.4 the application of random coefficients models is demonstrated, again using data from the repeated measures cross-over study taken from Jones and Kenward (1989).

8.2 Multi-centre trials with repeated measurements

Multi-centre trials with repeated measurements are considered in Section 8.2. A mixed model fitting centre and centre.treatment effects as random, and also using a covariance pattern to model correlations between the repeated measurements is specified. Mixed models with this structure are illustrated in Section 8.2.1 using data from the hypertension trial introduced in Section 1.3. In this example the results obtained were similar to the analysis ignoring the effects of centres carried out in Section 6.3. However, treatment effects estimates were slightly more accurate due to the inclusion of random centre effects.

8.3 Multi-centre cross-over designs

The multi-centre cross-over design is considered in Section 8.3. Three alternative mixed models are considered with different specifications for centre and centre.treatment effects. It is recommended that centre and patient effects are always fitted as random to allow any additional information on treatments from the patient and centre error strata to be recovered. If, additionally, it is assumed that treatment effects vary randomly across centres a wider inference can be obtained by fitting centre.treatment effects as random as described for the multi-centre design.

8.4 Multi-centre designs with centres grouped hierarchically

Multi-centre designs where centres are grouped hierarchically, for example by country, are considered in Section 8.4. A mixed model fitting centre, country, centre.treatment and country.treatment effects as random is suggested. This model assumes that treatment effects vary across centres and countries and leads to wider inferences. It also allows shrunken estimates of treatment effects at each centre and country to be calculated.

8.5 Bioequivalence studies with replicate cross-over designs

Bioequivalence studies with replicate cross-over designs are considered in Section 8.15. In these designs treatments are received more than once by each patient. They can be used to assess the within subject variability on each treatment, the variability of the within subject treatment difference, and the between subject variability of each treatment. The parameters obtained can then be used to establish 'individual bioequivalence' and 'population bioequivalence'. A common design has four periods and patients are randomised to receive treatments in one of the sequences: ABAB, ABBA, BABA, BAAB. Approaches based on analysing summary statistics calculated at each visit (eg AUC, time to maximum concentration) are considered for this design.

The published material considers several mixed model approaches with a variety of covariance structures. The implications of each model are considered in depth including those suggested by the FDA (the Food and Drug Administration). It is suggested that the model used to assess average bioequivalence should be chosen by using likelihood ratio tests to statistically justify the inclusion of variance and covariance terms in the model. The FDA guidance specifies that a sequence and sequence by period interaction are fitted as fixed effects. However, the inclusion of these effects has no bearing on the overall bioavailability results. Special consideration is given to approaches for establishing population and individual bioequivalence and a novel approach to calculating confidence intervals for the criteria using a Bayesian model is suggested.

A worked example illustrates several alternative models for analysing a four period trial to compare two formulations of an anti-anxiety agent where patients were randomised to receive treatments in the sequences ABAB, ABBA, BABA, BAAB. It was taken from the FDA website of datasets and utilises data provided by GlaxoSmithKline Pharmaceuticals (http://www.fda.gov/cder/bioequivdata). In this example the choice of covariance structure did not affect the conclusion in relation to average bioequivalence. However, the probability intervals resulting from the Bayesian model were wider than the confidence intervals obtained from the REML models indicating that the Kenward-Roger standard error adjustment may not have overcome the bias known to occur in fixed effects standard errors. The posterior sample obtained from a Bayesian analysis is used to calculate population and individual bioequivalence criteria and to provide exact probability intervals. In contrast to other complex approaches that have been suggested for calculating confidence intervals for the criteria, this method is relatively straightforward and the resulting probability intervals are unbiased.

9. Discussion

The material submitted in this thesis has provided an wide-ranging examination of mixed models for both normal and non-normal data for a variety of designs and data structures used in the medical field and beyond. This chapter reflects on the factors underlying the establishment of mixed models as a mainstream technique, considers some of the difficulties posed by the fitting methods, discusses the choice of models available, and highlights areas where more work is required.

9.1 Mixed models in perspective

The benefits of mixed models have been known about for some time, see for example Yates (1940), but it is only in recent years that their use has become widespread. It is interesting to consider the reasons behind this. One factor is undoubtedly the increased availability of computer power. When mixed models were first proposed carrying out a maximum likelihood analysis would have been time-consuming, however such an analysis is now virtually instant and computer power is not usually a consideration. Adequate computer power has encouraged software producers to include mixed models software into their packages. In 1989 mixed models software was introduced into two well known packages, Genstat and SAS, making the approach widely accessible to practising statisticians. The introduction of the comprehensive MIXED procedure into SAS had particular significance as this package was widely used within the pharmaceutical industry. However, software producers need to bear in mind the influence that their method of application and presentation of results will have on a user's understanding of a new methodology and interpretation of results. Unless the user's knowledge is comprehensive there may be a tendency to accept results produced by software regardless of whether or not they are completely appropriate for a situation. For example, early versions of PROC MIXED were misleading because the correct DF was not used for F and t tests and inappropriate z-tests were given for variance components. Thus, while the statistician always has ultimate responsibility for the integrity of an analysis, software developers need to assume some responsibility for ensuring an appropriate and clear presentation of results and for providing thorough documentation.

The ready availability of software and computer power are not by themselves sufficient to cause a methodology to become mainstream. There need to be clear improvements over other available methods, the new technology should not be so complex that it cannot be understood by the majority of practising statisticians and there needs to be an area of application where it may be frequently applied. Mixed models are one of several techniques suitable for the common situation involving an experiment where the effect of one or more factors is examined in relation to an outcome variable and where typically an objective is to test the null hypotheses that a particular factor (eg treatment) has no effect on the outcome variable. Previously methods such as t-tests, ANOVA and regression along with other analogous methods for non-normal data have been the mainstay for analysing this type of experiment. Mixed models are important because they too are suitable for analysing such experiments and additionally offer the advantage of taking into account correlations in the data. Although more complex, the broad principles can be understood by the majority of statisticians. By contrast, many other new and complex techniques which have become available with increased computer power have not found a place in the average statistician's repertoire because they are less easily understood, despite offering advantages for particular situations.

Mixed models have also been developed in parallel in the social sciences where they are referred to as 'multilevel' models. However, there has been little interaction between the alternative disciplines during these developments and different terminologies and techniques have evolved. Despite this, the main features of both approaches to modelling are identical and the same results are usually obtained by the different fitting techniques. Recently there has been more recognition of the overlap between the two areas of application and there have been examples of the MIXED procedure in SAS being used to fit 'multilevel' models by social scientists. There would appear to be potential for a more unified development in the future, particularly as issues associated with the approach are common to both areas.

9.2 Fitting methods

A variety of methods are available for fitting mixed models. Techniques based on optimisation such as maximum likelihood or generalised iterative least squares have been most frequently applied. However, the use of these techniques has several consequences. Negative variance components can be obtained even though they are not permitted by the definition of random effects which are assumed to have a distribution with a positive variance. Standard error estimates for fixed and random effects are biased downwards and although approximations are available to correct for this bias (see Section 2.4.3), it is never clear whether the corrections are completely adequate. For example, in the bioequivalence study described in Section 8.15 the Bayesian standard error for the treatment difference was notably different from the standard error obtained using REML even after correction with the Kenward-Roger approximation. Degrees of freedom for significance tests also need to be computed using approximations to allow for the different error strata used for estimating effects from mixed models (see Section 2.4.4). Although the use of approximations will provide satisfactory results in most situations, it may cause us to consider whether optimisation methods form the most natural approach for fitting mixed models. More work is required to examine the adequacy of approximations for standard errors, degrees of freedom and the resulting significance tests produced, across a variety of experimental situations.

The Bayesian approach, however, is not based on using optimisation and the problems described above are therefore not encountered. Rather than optimising a criterion, the joint distribution of the model parameters is obtained and this allows exact standard errors, probability intervals and 'Bayesian p-values' to be derived. Its use, with uninformative priors, is therefore perhaps the more suited for fitting mixed models as the use of approximations is avoided. It also allows more complex statistics from an analysis such as measures of population and individual bioequivalence and their confidence intervals (see Section 8.15) to be derived. However, its use is far less widespread and accepted than optimisation approaches. There are likely to be several reasons for this. There is a lack of understanding of the approach by statisticians whose training is usually based on frequentist techniques and who may sometimes assume that Bayesian methods necessarily incorporate prior information. A lack of user-friendly software has also made these methods less accessible to statisticians. Although a Bayesian analysis can be carried out in PROC MIXED, it requires several manual steps which would not be immediately apparent to statisticians who do not already have a firm understanding of the approach (see Section 2.3). While packages such as BUGS are freely available for carrying out Bayesian analyses, they are not as user friendly as well established packages such as SAS and appear to be little used by statisticians outside academia. A potential drawback of the Bayesian approach is that simulation methods are frequently used and the user is left to determine when the posterior distribution has been adequately sampled. More systematic approaches for assessing this are needed to allow such analyses to be applied reliably in clinical trials. This is an area where further work is required. Lastly there can be a reluctance by regulatory bodies and medical journals to accept results from a Bayesian analysis. This is only likely to change if the approach becomes more widely understood and systematic methods to determine when the posterior is adequately sampled become available.

Mixed model methods for analysing non-normal data based on optimisation (eg pseudo-likelihood, generalised estimating equations) are more complex than those for normal data. Usually a two-stage optimisation approach is required and approximations are used within the optimisation process as well as for calculating standard errors and degrees of freedom. A second complexity with both GLMMs and mixed models for categorical data is the discrepancy between the expected means and variances which occurs because, in the common situation where a 'one parameter' binomial or Poisson distribution is assumed, the expected variances are computed from the expected means based on shrunken random effects estimates (see Section 3.3.5). This discrepancy occurs regardless of whether the models are fitted using a technique based on optimisation or a Bayesian approach. Although the use of a dispersion parameter helps to overcome the biases caused, it is not completely adequate and it can be unclear when a satisfactory analysis has resulted. GLMMs are an area where more research is urgently required, particularly with the introduction of user-friendly software into statistical packages such as PROC GLIMMIX in SAS.

9.3 Choice of model

The published material describes several potential advantages offered by mixed models compared to more conventional fixed effects methods. These may relate to a gain in efficiency, an alternative interpretation, a more appropriate model, or a combination of these. However, careful consideration is required in deciding whether a mixed model should be used and, if so, which type of mixed model is the most appropriate.

In some situations the use of a mixed model may lead to more accurate estimates of treatment effects, for example when patient effects are fitted as random in a cross-over trial. However, this is not the case if the data are balanced and sometimes the gains in efficiency may disappear once approximations have been applied to correct bias in the standard error estimates.

In other situations, the use of a mixed model may give rise to potential new inferences. For example, in multi-centre trials treatment effects can be assumed to vary randomly across the centres by fitting centre.treatment effects as random. This gives a wider treatment effect standard error than in a fixed effects model and a wider generalisation of effectiveness over the population of centres can be made. Thus a mixed model here presents the analyst with a choice of model relating to the inference they would like to make. However, it is important that the limitations of alternative inferences are understood. For example, the representiveness of the centres included in a multi-centre trial must be taken into account since centres are rarely sampled at random.

For repeated measures data with missing values, an ANOVA or regression approach is inappropriate for modelling treatment effects because the correlations between the repeated measurements are not adequately accounted for. Here a mixed model provides a clearly preferable approach with its ability to model correlations between observations. Additionally it also provides an attractive alternative to the 'last value carried forward' approach when the primary objective is to analyse the last repeated measurement. There is a wide choice of covariance patterns available and it is possible to search for the pattern that is most appropriate to the covariance structure of the data. However, it should be borne in mind that the choice of pattern may affect results, particularly when data are unbalanced. In an experimental situation, an open-ended facility to search for a pattern is potentially open to misuse, particularly when results are to be submitted to a regulatory body. Thus it is desirable when writing statistical analysis plans to specify both a strategy for selecting a pattern, and also the range of patterns that will be considered. However, at present it is unclear which search strategy or range of covariance patterns are appropriate for a specific situation. This is another area where further work is required.

9.4 Summing up

Mixed models are unarguably more complex and involve more assumptions than most fixed effects models and thus more expertise is required to use them effectively and more explanation is needed when presenting results to non-statisticians. Looking more broadly at the use of modelling, situations where a single model or technique forms the only appropriate approach for an analysis are unusual and often several alternative models will adequately represent the 'truth'. However, from the opposite perspective there is rarely a technique that is perfect for a situation, for example there is unlikely to ever be complete certainty that the normality assumption for residuals is satisfied. Thus the selection of a model always involves a degree of compromise. It is also necessary to take into account other factors such as its complexity and the ease of interpretation of results when deciding which approach to use. Sometimes a simpler technique that is readily understood may be preferable to a more complex one, even when the latter offers advantages. For example, in Section 6.2 of the review a situation is described where GLMM analyses were criticised by British Medical Journal reviewers due to not being mainstream, and the less efficient, but better known, conditional logistic regression approach was suggested.

Considering the complexity as a factor influencing model choice brings to the fore the question of when it is appropriate for non-statisticians to perform statistical analyses. Ideally statistical analysis would always be performed by an analyst who has a full understanding of the methodology. However, there are many practical reasons why it is often desirable for researchers who are not professional statisticians to carry out their own analyses and the availability of user-friendly software has made this feasible. This may sometimes be satisfactory when the broad principles of a modelling approach can be understood by a non-professional statistician but less so when an approach is more complex. For example, checking normality and producing summaries and test results for a simple comparison of groups (eg t-tests and regression analysis) is relatively straightforward and can, with appropriate support, often be satisfactorily carried out by a non-professional statistician. However, this is less advisable when a methodology is more complex and mixed models would appear to fall into this category. The use of statistical analysis by non-professional statisticians is perhaps an area where the statistical profession need to establish a clearer view?

In summary, although mixed models are more complex than fixed effects approaches, this drawback is frequently outweighed by their advantages and their use is becoming more widespread as these are recognised. However, it is important that the growth takes place at a pace that allows issues requiring further examination to emerge and for these to be examined by researchers. This will provide practising statisticians with access to sufficient knowledge to allow them to use the approach with confidence.

Bibliography

Alexander F, Roberts MM, Lutz W, Hepburn W. (1989) Randomisation by cluster and the problem of social class bias. J Epidemiol Commun Health, 43:29-36.

BUGS (Bayesian use of Gibbs Sampler), www.mrc-bsu.cam.ac.uk/bugs, MRC Biostatistics Unit, Cambridge

Genstat, http://www.vsn-intl.com/genstat/, VSN International Ltd, Hemel Hempstead.

Hall, S, Prescott, RJ, Hallam, RJ, Dixon, S, Harvey, RE and Ball, SG (1991) A comparative study of Carvedilol, slow release Nifedipine and Atenolol in the management of essential hypertension', *Journal of Pharmacology*, **18**(4) S36–S38.

Hunter, KR, Stern, GM, Laurence, DT and Armitage, P (1970) Amantadine in Parkinsonism, *Lancet*, **i**, 1127–1129.

Jones, B and Kenward, MG (1989) *The Analysis of Cross-over Trials*, Chapman & Hall, London.

Kenward, MG and Roger, JH (1997) Small sample inference for fixed effects from restricted maximum likelihood, *Biometrics*, **53**, 983–997.

Lipsitz, SR, Kyungmann, K and Zhao, L (1994) Analysis of repeated categorical data using generalized estimating equations' *Statistics in Medicine*, **13**, 1149–1163.

Mead, R (1988) *The Design of Experiments*, Cambridge University Press, Cambridge.

Prescott, RJ (1981) The comparison of success rates in cross-over trials in the presence of an order effect, *Journal of the Royal Statistical Society, Series C*, **30**, 9–15.

Roberts, MM, Alexander, FE, Anderson, TJ and 9 others (1990) Edinburgh trial of screening for breast cancer: mortality at seven years. Lancet, 335, 241-246.

Roberts, MM, Alexander, FE, Anderson, TJ and 7 others (1984) The Edinburgh trial of screening for breast cancer: description of method. Br. J. Cancer, 47, 1-6

SAS/STAT documentation SAS/STAT 9.1 User's Guide, http://support.sas.com/documentation/onlinedoc/sas9doc.html

SAS PROC GLIMMIX documentation, http://support.sas.com/rnd/app/papers/glimmix.pdf

Senn, S (2002) *Cross-over Trials in Clinical Research*, second edition, John Wiley &

Sons, Chichester.

Talbot, M (1984) Yield variability of crop varieties in the UK, *Journal of Agricultural Science, Cambridge*, **102**, 315–321.

Thall, PF and Vail, SC (1990) Some covariance models for longitudinal count data with overdispersion, *Biometrics*, **46**, 657–671.

Thrusfield, M (1995) *Veterinary Epidemiology*, Blackwell Science, Oxford.

Yates, F (1940) The recovery of inter-block information in balanced incomplete block designs, *Annals of Eugenics*, **10**, 317–325.

Research publications

The candidate's contribution to research publications is coded as:-

(a) primarily the work of the candidate
(b) equal contribution by the candidate and other authors, or where the candidate is responsible only for statistical analysis and interpretation

1. Brown,H.K., Prescott,R.J. *Applied mixed models in medicine*, second edition, John Wiley, Chichester, 2006. (a, except where indicated in review)

1a. Brown,H.K., Prescott,R.J. *Applied mixed models in medicine*, John Wiley, Chichester, 1999. (a, except where indicated in review)

2. Grigor,P.N., Cockram,M.S., Steele,W.B., Le Sueur,C.J., Forsyth,R.E., Guthrie,J.A., Johnson,A.K., Sandilands,V., Reid,H.W., Sinclair,C. and **Brown,H.K.** 2001. Effects of space allowance during transport and duration of mid-journey lairage period on the physiological, behavioural and immunological responses of young calves during and after transport. Animal Science, 2001, 73:341-360. (b)

3. Alexander,F.E., Anderson,T.J., **Brown,H.K.**, Forrest,A.P.M., Hepburn,W., Kirkpatrick,A.E., Muir,B.B., Prescott,R.J., Smith,A., Warner,J. 'The Edinburgh randomized trial of breast-cancer screening - results after 14 years of follow-up', 1999, Lancet (b)

4. Alexander,F.E., **Brown,H.K.**, Prescott,R.J. 'Improved classification of socio-economic status explains differences in all-cause mortality in a randomised trial of breast cancer screening, 1998, *Journal of Epidemiology and Biostatistics*, 1998, 3, 219-224 (b)

5. Luis-Fuentes,V., Moran,L., Schber,K., Dukes-McEwan,J., **Brown,H.**, Sutherland,G.R., McDicken,W.N. 'Measurement of cyclic variation in ultrasonic integrated backscatter in continuously unsedated clinically normal dogs', *American Journal of Veterinary Research*, 1997, 58, 1055-59 (b)

6. Brooke,H., Gibson,A., Tappin,D., **Brown,H.** 'Case-control study of sudden infant death syndrome in Scotland, 1992-1995', *British Medical Journal*, 1997, Vol.314, 1516-1520 (b)

7. Dixon,J.M., Ravisekar,O., Cunningham,M., Anderson,E.D.C, Anderson,T.J., **Brown,H.K.** 'Factors affecting outcome of patients with impalpable breast cancers detected by breast screening', *British Journal of Surgery*, 1996, 83, 997-1001 (b)

8. Dallak,M., Luan,X.J., **Brown,H.,** Pirie,L. 'Stability of breathing patterns in men, rats and rabbits' *Journal of Physiology - London*, 1995, 483, 96-97 (b)

9. Humphreys,F., Symons,J.A., **Brown,H.K.**, Duff,G.W., Hunter,J.A.A. 'The effects of gamolenic acid on adult atopic eczema and premenstrual exacerbation of eczema', *European Journal of Dermatology*, 1994, 4, 598-603, (b)

10. Brown,H.K., Kempton,R.A., 'The application of REML in clinical trials', *Statistics in Medicine*, 1994, 13, 1601-1617, (a)

11. Alexander,F.E., Anderson,T.J., **Brown,H.K.**, Forrest,A.P.M., Hepburn,W., Kirkpatrick,A.E., McDonald,C., Muir,B.B., Prescott,R.J., Shepherd,S.M., Smith,A., Warner,J. 'The Edinburgh randomised trial of breast-cancer screening - results after 10 years of follow-up', *British Journal of Cancer*, 1994, 70, 542-548 (b)

Animal Science 2001, 73: 341-360
© 2001 British Society of Animal Science

1357-7298 / 01 / 09050341$20·00

Effects of space allowance during transport and duration of mid-journey lairage period on the physiological, behavioural and immunological responses of young calves during and after transport

P. N. Grigor[1], M. S. Cockram[1]†, W. B. Steele[1], C. J. Le Sueur[1], R. E. Forsyth[1], J. A. Guthrie[2], A. K. Johnson[2], V. Sandilands[2], H. W. Reid[3], C. Sinclair[3] and H. K. Brown[4]

[1]*Department of Veterinary Clinical Studies, University of Edinburgh, Easter Bush Veterinary Centre, Easter Bush, Roslin, Midlothian EH25 9RG, UK*
[2]*Institute of Ecology and Resource Management, University of Edinburgh, Agriculture Building, West Mains Road, Edinburgh EH9 3JG, UK*
[3]*Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Midlothian EH26 0PZ, UK*
[4]*Department of Public Health Sciences, University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK*

† Corresponding author e-mail: M.S.Cockram@ed.ac.uk

Abstract

The effects of space allowance during transportation and duration of a mid-journey lairage period on measurements of stress, injury, dehydration, food restriction and rest in young calves were assessed during and after transport. Groups of calves were transported for two 9-h journeys (at a space allowance of either 0·375 or 0·475 m^2 per calf) separated by a mid-journey lairage period of either 1 or 12 h. Non-transported calves were offered milk replacer and drinking water either at the usual times or only at the same times as the transported calves.

During transport, transported calves spent significantly less time lying down and had a greater plasma cortisol concentration than control calves. Under the driving conditions used, increased space allowance was not associated with greater injury or loss of stability. The duration of the mid-journey lairage was not an important factor; the shorter lairage time, giving the calves sufficient time to receive milk replacer but little opportunity to rest, had no major detrimental effects on the variables used to assess welfare. Although there was little evidence that transport affected immunological variables, there was some evidence that it adversely affected the health of the calves post transport.

Keywords: *animal welfare, behaviour, calves, physiology, transport.*

Introduction

When young calves are transported, they are potentially exposed to a number of factors that, either on their own or in combination, could affect their welfare (Cockram and Mitchell, 1999). Many aspects of the transportation of young calves are regulated by legislation. In the UK, the Welfare of Animals (Transport) Order 1997 (Great Britain Parliament, 1997) implements European Directive 91/628/EEC (European Council, 1991) as amended by Directive 95/29/EC (European Council, 1995) and depending on the type of vehicle used, this legislation stipulates: the journey length and structure, the intervals at which the calves should be offered food, water and rest, and provides recommended space allowances to be used during transport. However, there have been few detailed studies to evaluate the scientific basis of the legislation and to identify the important factors likely to affect the welfare of young calves during transport. The effects of transport on the welfare of cattle and calves have been reviewed by Tarrant (1990), Trunkfield and Broom (1990), and Knowles (1999).

Young calves can be stressed (Johnston and Buckland, 1976; Kent and Ewbank, 1986), dehydrated, fatigued (Mormede *et al.*, 1982; Atkinson, 1992) and experience either muscular exertion or damage (Knowles *et al.*, 1997 and 1999) as a result of transport. Transportation of young calves has also been associated with an increased risk of mortality from diseases, such as pneumonia that occur several weeks post transport (Staples and Haugse, 1974; Knowles, 1995). Prolonged transport, and inadequate and irregular feeding and watering are considered to be risk factors for 'shipping fever' in that they reduce the resistance of cattle sufficiently for viruses, such as bovine herpesvirus type 1 (BHV-1) to infect the respiratory tract, further reduce the animal's resistance and allow bacteria, such as *Pasteurella haemolytica*, to spread and cause fibrinous pneumonia (Yates, 1982; Roth and Kaeberle, 1982). Immunosuppression has been reported in weaned calves following either transportation (Blecha *et al.*, 1984; Murata, 1989; Murata and Hirose, 1990 and 1991) or administration of ACTH (Roth *et al.*, 1982). In the present study, the measurement of specific immunological responses to BHV-1 inoculation was used as a potential means of assessing whether various transport conditions were sufficient stressors to induce clinically relevant immunosuppression.

This paper examines the effects on the welfare of calves of a journey structure that represents the maximum times that young calves could be transported (i.e. two 9-h journeys separated by a mid-journey resting or lairage period), compares the effects of the minimum mid-journey lairage duration (1 h) with an extended lairage duration of 12 h, and examines the effects of transporting calves at the extremes of the space allowances recommended in current legislation. In particular, the experiment investigated whether increasing the duration of the mid-journey lairage provided greater opportunity for rest and to replenish nutrient energy used and water lost during the previous journey. The extremes of space allowances currently recommended for the transport of young calves were compared to investigate whether the higher space allowance provided a greater opportunity for the calves to lie down and adjust their posture during a long journey without reducing their stability and increasing their risk of stress and injury.

Material and methods

Animals and management

Over a period of 6 months, four batches of 24 unweaned, 10-day-old male Holstein-Friesian calves were transported for about 1 h from a livestock market to the experimental accommodation. The calves were housed in single pens (1·95 m × 1·24 m) with straw bedding (replenished at intervals to ensure a comfortable and dry lying area) and *ad libitum* water from a bucket. On the day of arrival, each calf was offered 2 l of an oral rehydration solution (Energaid, Elanco, Co. Down) and given an intramuscular injection of oxytetracycline (Engemycin LA, Intervet, Cambridge, UK). The calves were offered 2 l of milk replacer (Denkavit Herdbuild 297, milk-based protein 210, oil 170, fibre 2·5 and ash 100 g/kg) at 38°C by bucket twice daily (08:00 and 15:00 h). After 17 days, the calves were also offered 200 g/day of concentrates (Dalgety Agriculture Ltd, 2222 Northern County Coarse Mixture, protein 180, oil 37·5, fibre 70 and ash 80 g/kg). After a total of 18 to 21 days (depending on initial consumption of concentrates), 300 g/day of concentrates were offered, increasing to 700 g/day after a total of 21 to 25 days.

Treatments

The effects of two 9-h periods of either transport (at a space allowance of either 0·375 m^2 per calf or 0·475 m^2 per calf) or food and water deprivation, separated by either a 1-h or a 12-h mid-journey lairage period were investigated. Within each of the batches, the calves were randomly allocated to one of four treatment groups (six calves per group), balanced by initial live weight, and initial serum concentrations of total gamma-globulin and BHV-1 antibody, as follows: non-transported fed controls (offered food and water during the transport periods); non-transported unfed controls (not offered food and water during the transport periods); transported at a space allowance of 0·375 m^2 per calf; and transported at a space allowance of 0·475 m^2 per calf. To minimize disturbance to the non-transported calves, these calves were allocated to a section of the accommodation separated from the section containing the transported calves by a central feeding area. The effects of transport, food and water deprivation and space allowance during transport were investigated within each batch but the effects of different lairage durations were investigated in different batches in the following order: 1-h, 12-h, 1-h and 12-h lairage durations. The effects of lairage duration were compared by starting the second journey at the equivalent time of day for all batches.

The mean initial live weights (measured on arrival at the unit) in the four batches of calves were 48 kg (s.e. 1·3), 49 kg (s.e. 0·9), 48 kg (s.e. 1·0) and 48 kg (s.e. 0·7), respectively, with the mean per treatment group ranging from 47 kg (s.e. 1·9) to 51 kg (s.e. 2·3). The mean initial serum concentrations (g/l) of total gamma-globulin in the four batches of calves were 7·6 (s.e. 1·00), 7·3 (s.e. 0·84), 8·2 (s.e. 0·80) and 8·23 (s.e. 0·92), respectively, with the mean value per treatment group ranging from 5·6 (s.e. 1·53) g/l to 8·7 (s.e. 2·4) g/l. The mean initial titres of BHV-1 antibody in the four batches of calves were 165 (s.e. 45·6), 232 (s.e. 109·0), 192 (s.e. 63·1) and 178 (s.e. 55·4) arbitrary units, respectively, with the mean value per treatment group ranging from 33 (s.e. 23·4) to 477 (s.e. 359·0) arbitrary units.

After each batch of calves had been housed for 8 days, the groups to be transported were loaded at either 12:00 h (batches with 12-h lairage duration) or 23:00 h (batches with 1-h lairage duration) on to a single deck, non-articulated vehicle and transported for 9 h (journey 1). The vehicle (described by Cockram *et al.*, 1996) consisted of an observation area and a livestock area divided into two pens by a weld mesh partition, a metal tread-plate floor and metal walls. The floor was covered with straw bedding that was replenished after each journey. In the first and fourth batches, the calves in the front pen (1·01 m long × 2·22 m wide) were transported at a space allowance of 0·375 m^2 per calf and those in the back pen (1·30 m long × 2·22 m wide) were transported at a space allowance of 0·475 m^2 per calf. In the second and third batches, the sizes of the pens were adjusted so that the calves transported at the 0·475 m^2 per calf space allowance were in the front pen and those transported at the 0·375 m^2 per calf space allowance were in the back pen. Each 9-h journey consisted of five similar circuits of about 80 min on public roads at an average speed of 50 km/h. The vehicle was stationary for a period of 25 min after each circuit with the engine switched off to allow blood sampling of the calves. Each journey consisted of proportionately 0·80 A roads (single and dual carriageway), 0·15 town roads and 0·05 unclassified roads. The median frequencies of occurrence of roundabouts, sharp corners, and starts/stops during a circuit were 14, 10 and 6·5, respectively.

When the transported calves had been loaded, the drinking water was removed from the non-transported, unfed group and no food was offered to these calves during the time that the transported calves were on the vehicle. The non-transported fed controls had access to drinking water during the time that the transported calves were on the vehicle and were offered food at the normal feeding time.

After journey 1, the transported calves were unloaded and returned to their pens for a lairage period of either 1 h or 12 h. Drinking water was offered to all groups at the start of the lairage period. During the 12-h lairage period, transported groups and non-transported unfed controls were offered milk replacer at the start of the lairage period, and all groups in all batches were offered milk replacer 1 h before the start of journey 2.

After the lairage period, the transported groups were reloaded into the same pens on the vehicle and transported for a further 9 h (journey 2), as described for journey 1. During journey 2, the non-transported calves were treated as described above for journey 1. After journey 2, the transported calves were unloaded and returned to their pens. Drinking water and milk replacer were offered to the non-transported, unfed group and the transported groups.

The transport treatment therefore involved several changes to the calves, including removal from their single pen, loading, group penning at one of two space allowances, withdrawal of milk replacer and drinking water, transportation, unloading and return to their original single pen.

Temperature and relative humidity

Air temperature and relative humidity were recorded at 5-min intervals using Tinytalk Data Recorders (Orion Components, Chichester). In the four batches, the mean air temperatures in the calf pens during observation periods were 17°C, 20°C, 15°C and 11°C, respectively, and in the vehicle during journeys were 18°C, 21°C, 16°C and 15°C, respectively. In the four batches, the mean relative humidities in the calf pens were 0·53, 0·61, 0·73 and 0·73, respectively, and in the vehicle were 0·53, 0·65, 0·73 and 0·61, respectively.

Blood biochemistry

Two days before the start of the transport treatments, a jugular cannula was inserted into each calf under local anaesthetic. An extension tube containing heparinized saline was connected to the cannula and taped to the dorsal area of the neck to allow manual blood sampling to be performed *via* a three-way tap with either minimal or no restraint of the calves. Seven ml of blood was manually collected from all treatment groups into 'Sarstedt' monovette tubes containing lithium heparin and 1 ml was decanted into a tube containing fluoride at the following times: at 3-h intervals for 24 h before the start of treatment, immediately before and after the transported calves were loaded (both journeys), at 1·75-h intervals during each journey, immediately after unloading and before feeding (both journeys), at 3-h intervals during the 12-h lairage (second and fourth batches), and at 3-h intervals for 24 h after journey 2 (post treatment). The jugular cannulae were removed 24 h after the end of journey 2.

The blood samples were refrigerated, and packed cell volume was measured using a Wifug haemicrofuge centrifuge. The blood samples were then centrifuged, and the plasma removed and stored at −20°C for subsequent analysis. The plasma osmolality was measured by freezing point depression using a micro-osmometer (Advanced Micro-Osmometer

Model 3 MO, Vitech Scientific Ltd, West Sussex). Plasma cortisol concentration was measured using a chemiluminescent enzyme immunoassay (Immulite Cortisol Kit LK CO5) on an Immulite Automated Analyser (Euro/DPC Ltd, Glyn Rhonwy, Llanberis, Gwynedd). Bovine standard serum (Multi Sera Elevated Randox Laboratories Ltd, Antrim) was assayed and was found to provide cortisol values similar to the quoted range for a radioimmunoassay (RIA) method (Immulite method, 513 nmol/l, RIA method, 303 to 511 nmol/l). Parallelism was demonstrated by serial dilution of the bovine standard serum (513, 251, 127, 63, 37 nmol/l respectively). The detection limit was 5 nmol/l. The inter- and intra-assay coefficients of variation were 0·04 and 0·16, respectively. The plasma creatine kinase activity (Bayer Diagnostics Kit T01-1882-01) and the plasma concentrations of free fatty acids (Randox Laboratories Kit FA/115S), glucose (Randox Laboratories Kit GL 586) and albumin (Bayer Diagnostics Kit 01137702) were measured on a Bayer Diagnostics RA-2000 random access chemistry analyser (Bayer Diagnostics, Basingstoke) at 37°C. The plasma concentrations of sodium, potassium and chloride were measured on a Corning 644 electrolyte analyser using ion-selective reagents.

Behaviour

The behaviour of all calves was recorded using Psion hand-held computers using the Observer Software (Noldus Information Technology, 1993 and 1994) for 24 h before the first journey, during both journeys, during the lairage period, and for 24 h after transport. Behaviour in the pens was monitored remotely *via* a video-monitoring system connected to a Sprite DX Multiplexer (Dedicated Micros, Manchester). The behaviour of calves on the transporter was recorded by direct observation from the semi-concealed area on the vehicle.

The calves were scan-sampled every 10 min and the occurrence of the following behaviour was recorded: standing-still: either alert (head raised, ears pricked) or non-alert (not involved in any of the following activities and not performing any other obvious activity); moving; lying; eating straw; drinking milk substitute; drinking water; oral investigation (licking or sniffing the pen walls); ruminating; self-grooming and interacting with another calf. The proportion of observations in which each behaviour was observed was calculated for each 3-h observation period during the 24-h pre- and post-transport periods and during the 12-h lairage period, for the 1-h lairage period, and for each of the circuits in journeys 1 and 2. During transportation, orientation (the direction in which an individual was aligned in relation to the direction of travel) was recorded during each scan. Except during blood-sampling, the behaviour of transported calves was also continuously observed during both journeys, and the following events were recorded: standing up, lying down, losing balance (momentary loss of body stability in which calf moves its foot position to remain standing), falling, colliding with vehicle wall, colliding with another calf, being trampled on by another calf; and expressed as the number of events per calf per hour. After unloading, the latencies of transported calves to drink water and to lie down were recorded. The latency of transported calves and non-transported non-fed calves to drink milk replacer following both journeys, and the number of calves requiring assistance to drink milk replacer, were also recorded.

Heart rate

Two days before the start of the transport treatments, each calf was fitted with heart rate recording equipment (Polar Accurex Plus, Polar Electro Oy, Kempele, Finland). The area behind the left foreleg was shaved, gauze and electrode gel were applied to each transmitter to aid conductivity and the transmitters and receivers were held in place by an elastic strap. Heart rate was recorded at 60-s intervals, beginning 24 h before the start of journey 1 until approximately 12 h after journey 2.

Live weight, food and water intakes

The calves were weighed on the day of arrival at the experimental accommodation, immediately before transport, immediately after journey 2, and at 1, 2 and 3 weeks post transport. Daily intakes of milk replacer, drinking water and concentrates were recorded for the duration of the experiment.

Immunological measurements

Serum total gamma-globulin concentration. On arrival at the unit, a blood sample was taken by jugular venipuncture and the total serum gamma-globulin concentration in each calf was determined using a tubometric method (Wolfson *et al.*, 1948).

BHV-1 infectivity in nasal swabs. On the day before transport, all calves were intranasally inoculated with the Oxford type strain of BHV-1 at a dose of approximately 1 × 10^6 plaque forming units (p.f.u.) per calf. Nasal swabs were taken immediately before inoculation and then daily for at least 7 days after inoculation. The weight of material adhering to each swab was recorded and the swabs stored in 5 ml of transport medium (1 × Hanks [Gibco BRL], 10 mg/ml BSA, 0·4 mg/ml sodium bicarbonate ($NaHCO_3$), 800 units/ml penicillin, 1 mg/ml streptomycin and 60 units/ml polymixin) at –70°C. Embryonic bovine tracheal cells (EBTr) were plated onto a 12 well tissue culture plate (Corning) (1 × 105 cell per ml, 1 ml per well) and grown overnight at 37°C in proportionately 0·05 carbon dioxide containing medium 1 × 199 (Gibco BRL) supplemented with proportionately 0·1 foetal calf serum, proportionately 0·1 tryptose phosphate broth, 100 units per ml penicillin, 0·1 mg/ml streptomycin, proportionately 0·01 amphotericin, proportionately 0·008 $NaHCO_3$ and 0·1 mol/l glutamine. The medium was removed from the cells and dilutions of the nasal swab sample added. The plates were incubated at 37°C in proportionately 0·05 carbon dioxide for 1 h and overlaid with media containing carboxymethyl cellulose. This medium was similar to the growth medium except that the concentration of foetal calf serum was proportionately 0·02 and carboxymethyl cellulose was added to a final concentration of proportionately 0·01. The plates were then incubated at 37°C in proportionately 0·05 carbon dioxide for a further 3 days and stained with a solution of proportionately 0·001 crystal violet and 0·1 methanol. Plaques due to virus infection were then counted and the results presented in p.f.u. per g of material adhering to the swab.

Lymphocyte stimulation assay. Ten ml heparinized blood samples were obtained by jugular venipuncture immediately before intranasal inoculation with BHV-1, and 7 and 14 days after inoculation. Lymphocyte stimulation assays were performed as described by Burrells and Wells (1977) and Burrells *et al.* (1995) using the following antigens: Concanavalin A (Con A) (ICN Biomedical, High Wycombe, UK) at a final concentration of 7·5 mg/ml as a positive control and inactivated BHV-1 (Psoralin, HRI associates) as the test antigen. The concentration of virus used for the test was 4 × 107 p.f.u. per ml (pre-inactivation).

Serum BHV-1 antibody assay. Seven ml blood samples were obtained by jugular venipuncture on the day that the calves arrived at the unit, immediately before intranasal inoculation with BHV-1 and at 5-day intervals over a 20-day period. The clotted samples were centrifuged and the serum stored at −20°C. The concentration of serum BHV-1 antibody was determined by an enzyme linked immunosorbant assay (ELISA) method as previously described by Lyaku *et al.* (1990). The plates were read using a Dynatech plate reader and results presented as arbitrary units related to the optical density of 490 mm of the sample read against concentration of a known positive serum on a standard curve.

Serum ovalbumin antibody concentration. All calves were given a 1 ml intra-muscular injection of 5 mg ovalbumin in Freund's incomplete adjuvant on the day before transport. Serum ovalbumin antibody measured 5, 10, 15, and 20 days after was titrated by using plates (Dynatech 129A) coated with ovalbumin at a concentration of 10 mg/ml in coating buffer (0·2 mmol/l $NaHCO_3$, 0·2 mmol/l Na_2CO_3).

Health

Rectal temperatures and clinical signs of disease were recorded daily for the duration of the experiment. Dehydration was scored on the basis of skin elasticity and presence of sunken eyes (Constable *et al.*, 1998). Diarrhoea was assessed by scoring faecal consistency. Respiratory signs were assessed by scoring respiration rate, ocular and nasal discharges, inflammation and lesions of the nasal mucous membranes and conjunctiva and presence of cough and salivation. Veterinary treatment was given when appropriate and treatments were recorded. The calves were euthanased 20 days post treatment and any gross pathological changes in the lungs were recorded.

Statistical analyses

The data were grouped into the following experimental periods: pre-treatment, journey 1, lairage, journey 2, and post treatment. For the scan-sampled behavioural data, heart rate data and blood variables, a repeated measures analysis of covariance (Laird and Ware, 1982) using the mixed procedure within SAS version 6 (SAS Institute Inc., Cary, USA) was used to examine the effects of 'period' (journey 1, lairage, journey 2, post treatment), 'Treatment' during transport periods (non-transported fed controls, non-transported unfed controls, transported at 0·375 m^2 per calf space allowance and transported at 0·475 m^2 per calf space allowance), and 'lairage' duration (1-h or 12-h). In addition to the main effects, the following interactions were examined: treatment × period, lairage × treatment, and lairage × treatment × period. As the responses of the calves may have been correlated within batches, batch effects were fitted in each model as random. Thus, all treatment and lairage effects were compared against a background of between-batch variation. To allow for between-treatment differences prior to the start of the transport period, pre-treatment values were used as a covariate, with analyses carried out on differences between pre-treatment values and means of values in subsequent periods. Where appropriate, data were log-transformed to maximize normality. These analyses allowed the following specific comparisons to be made: (a) treatment differences during each period, (b) lairage differences and (c) differences between the first and second journeys. These follow-up analyses produced many pair-wise comparisons. To avoid the problems of multiple testing, a Bonferroni adjustment (in which the *P* values are multiplied by the number of comparisons of the effect) was applied to the resultant P values. Where the main effects were non-significant (e.g. treatment × period), but the corresponding estimate comparisons produced significant effects, treatment effects are reported as being statistically significant, as these latter comparisons were of primary interest. Where it was considered to be appropriate to investigate treatment effects at the following specific time-points: (a) each circuit of both journeys and each 3-h post-treatment observation period for scan-sampled behavioural data and (b) each blood sample on both journeys and during the 24-h post-treatment period, the same repeated measures model was used, except that the observation period (or sample number) was substituted for the period effects. The same overall method was used to analyse live-weight change, water intake, immunological measurements and rectal temperature, except that the post-treatment period was divided into appropriate periods. The effect of space allowance during transport on the median frequency of continuously recorded events was examined during journey 1 using Mann-Whitney tests. The effects of lairage duration on the median frequency of events during journey 2 was analysed using Mann-Whitney tests. The effects of transport, food and water deprivation, space allowance during transport and lairage duration on the incidence of respiratory disease and the incidence of dehydration/diarrhoea during the post-treatment period were investigated as a stratified cohort study using Win Episcope 1. Spearman's rank correlations were performed to investigate relationships between some physiological, immunological and health measurements.

During the week before the start of the treatment period of the fourth batch, some calves had clinical signs of dehydration and diarrhoea. Four calves were replaced and two calves allocated to the fed control group were removed from the experiment.

Results

Statistical significance of main effects and interactions are shown in Table 1. Table 2 shows effects of transport, lairage duration during transport, space allowance during transport and food and water withdraw on mean values of blood chemistry, lying behaviour and water intake of young calves. Table 3 shows effects of transport, space allowance during transport and food and water withdrawal on further mean values of blood chemistry and live weight of young calves.

Table 1 *Statistical significance of main effects and interactions of lairage duration (1 h and 12 h) (L), treatment (space allowance during transport, non-transport with normal food and water, and no food and no water) during transport period (T), and experimental period (pre-treatment, journey 1, lairage, journey 2 and post-treatment) (P) on young calves*

	Significance					
	L	T	P	T×P	L×T	L× T×P
Plasma cortisol concentration		***	***	***		
Plasma creatine kinase activity		**	**	*	*	**
Packed cell volume		***	***			
Plasma albumin concentration						
Plasma osmolality		**			*	*
Plasma sodium concentration				*		
Plasma potassium concentration			***	*		***
Plasma chloride concentration			***			
Plasma glucose concentration			***	***		***
Plasma free fatty acid concentration		***	***	***		***
Lying down (mean proportion of scans)	***	***	***	***		***
Heart rate		*				***
Live-weight change (day 0 to day 20 post treatment)			***			
Water intake (day 1 post treatment)			***	**		**
BHV-1 infectivity in nasal swabs			***	*		
Lymphocyte stimulation assay						
Serum BHV-1 antibody titre			***		*	**
Serum ovalbumin antibody titre			***			
Rectal temperature		**	***			

Table 2 *Effects of transport, lairage duration during transport, space allowance during transport, and food and water withdrawal on mean values of blood chemistry, lying behaviour and water intake of young calves*

Lairage duration		1 h				12 h			
Treatment		No transport		Transport		No transport		Transport	
Treatment during transport period Variable	Experimental period	Normal feeding	No food and no water	Space allowance 0·375 m^2 per calf	Space allowance 0·475 m^2 per calf	Normal feeding	No food and no water	Space allowance 0·375 m^2 per calf	Space allowance 0·475 m^2 per calf
Plasma creatine kinase activity (IU/l)	Pre-treatment	180·6	167·6	167·6	152·3	170·8	149·1	161·1	146·6
	Journey 1	168·3^{y}	163·4	212·7	148·5	165·6^{y}	155·4	177·7	185·9
	Lairage	148·6^{a}	150·0	224·8^{b}	146·6^{b}	148·8^{a}	149·1	160·9^{b}	155·0^{b}
	Journey 2	146·3^{ax}	159·5	240·0^{b}	158·0^{b}	144·4^{ax}	145·4	144·0^{b}	138·1^{b}
	Post-treatment	148·6^{a}	164·3^{b}	207·3^{b}	169·7^{b}	138·7^{a}	156·1^{b}	134·9^{b}	136·3^{b}
Plasma osmolality (mosmol)	Pre-treatment	277·6	276·9	280·2	281·9	288·0	282·6	282·8	283·9
	Journey 1	276·4	273·8	281·7	281·0	286·5	282·0	283·7	281·9
	Lairage	277·8	274·0	282·2	281·0	282·6	283·0	283·3	281·4
	Journey 2	278·5	273·4	283·4	282·9	283·1	281·7	283·2	280·1
	Post-treatment	278·8	276·5	282·8	281·9	285·0	282·0	282·7	279·7
Plasma potassium concentration (mmol/l)	Pre-treatment	4·81	4·73	4·68	4·47	4·49	4·65	4·66	4·61
	Journey 1	4·84	4·87	4·77^{x}	4·69^{x}	4·54	4·61	4·42^{x}	4·47^{x}
	Lairage	4·77	4·78	4·76	4·81	4·46	4·68	4·67	4·70
	Journey 2	5·08	5·09	5·01^{y}	4·99^{y}	4·51	4·60	4·49^{y}	4·58^{y}
	Post-treatment	4·88	4·79	4·97	4·80	4·57	4·57	4·51	4·61
Plasma glucose concentration (mmol/l)	Pre-treatment	4·69	4·73	4·73	4·50	5·44	5·14	5·07	5·11
	Journey 1	4·52^{b}	4·40^{x}	4·19^{ax}	3·96^{ax}	5·20^{b}	4·70^{x}	4·60^{ax}	4·46^{ax}
	Lairage	4·08	4·48	3·93	3·80	4·75	4·92	5·15	4·90
	Journey 2	4·96	4·77^{y}	4·94^{y}	4·67^{y}	4·96	4·86^{y}	4·90^{y}	4·79^{y}
	Post-treatment	4·80	4·95	4·89	4·65	4·86	4·94	5·01	4·74
Plasma free fatty acid concentration (mmol/l)	Pre-treatment	0·31	0·28	0·23	0·27	0·25	0·26	0·21	0·21
	Journey 1	0·33^{a}	0·32^{ax}	0·42^{by}	0·39^{by}	0·13^{a}	0·25^{ay}	0·33^{by}	0·33^{by}
	Lairage	0·44^{b}	0·38^{aq}	0·44^{q}	0·38^{q}	0·37^{b}	0·24^{aq}	0·27^{p}	0·23^{p}
	Journey 2	0·25^{a}	0·22^{ax}	0·25^{bx}	0·27^{bx}	0·18^{a}	0·20^{ax}	0·26^{bx}	0·28^{bx}
	Post-treatment	0·25	0·21	0·24	0·24	0·29	0·22	0·23	0·25
Lying down (proportion of scans)	Pre-treatment	0·76	0·77	0·72	0·73	0·74	0·74	0·75	0·77
	Journey 1	0·83^{b}	0·88^{b}	0·34^{ax}	0·41^{ax}	0·71^{b}	0·62^{b}	0·22^{ax}	0·37^{ax}
	Lairage	0·17^{ap}	0·32^{p}	0·44^{bp}	0·32^{bp}	0·76^{aq}	0·82^{q}	0·87^{bq}	0·83^{bq}
	Journey 2	0·78^{b}	0·75^{b}	0·52^{ay}	0·54^{ay}	0·66^{b}	0·76^{b}	0·50^{by}	0·64^{by}
	Post-treatment	0·80	0·80	0·87	0·84	0·77	0·80	0·81	0·80
Water intake (l/day)	Pre-treatment	0·46	0·62	0·22	0·35	0·43	0·54	0·50	0·35
	Lairage	0·10	0·01	0·03	0·00	0·21	0·39	0·22	0·21
	Day 1 post treatment	0·14^{a}	0·60^{b}	0·18^{a}	0·25^{a}	0·25^{a}	0·53^{b}	0·39^{a}	0·29^{a}
No. of calves		12	12	12	12	10	12	12	12

[a,b] Different superscripts within a row indicate a significant difference between treatments within an experimental period ($P < 0·05$).

[p,q] Different superscripts within a row indicate a significant difference between lairage durations within a treatment ($P < 0·05$).

[x,y] Different superscripts within a column indicate a significant difference between journey 1 and journey 2 within a treatment ($P < 0·05$).

Plasma cortisol concentration

Treatment comparisons during experimental periods showed that during journey 1, plasma cortisol concentration was significantly greater in transported calves than in control calves ($P < 0·001$; Figure 1).

able 3 *Effects of transport, space allowance during transport, and food and water withdrawal on mean values of blood chemistry and ,e weight of young calves*

reatment		No transport		Transport	
reatment during ransport period			No food	Space allowance	Space allowance
ariable	Experimental period	Normal feeding	and no water	0·375 m^2 per calf	0·475 m^2 per calf
lasma cortisol concentration (nmol/l)	Pre-treatment	15·9	23·6	12·7	17·9
	Journey 1	12·5^{a}	16·3^{a}	25·0^{by}	25·2^{by}
	Lairage	13·5	15·6	10·8	17·9
	Journey 2	10·3	14·0	16·4^{x}	18·0^{x}
	Post treatment	11·0	12·4	13·1	13·6
acked cell volume	Pre-treatment	0·364	0·364	0·369	0·375
	Journey 1	0·354^{y}	0·349^{y}	0·358	0·364
	Lairage	0·354	0·348^{a}	0·364^{b}	0·374^{b}
	Journey 2	0·343^{x}	0·339^{ax}	0·354^{b}	0·363^{b}
	Post treatment	0·343	0·332^{a}	0·349^{b}	0·362^{b}
lasma albumin concentration (g/l)	Pre-treatment	31·7	32·6	32·2	31·5
	Journey 1	31·6	32·7	32·7	32·0
	Lairage	31·5	32·5	32·8	32·2
	Journey 2	32·3	32·3	33·3	32·0
	Post treatment	32·0	32·1	32·2	31·9
lasma sodium concentration (mmol/l)	Pre-treatment	136·0	134·1	136·2	135·5
	Journey 1	135·3	133·5	136·7	135·9
	Lairage	135·1	133·7	136·5	135·8
	Journey 2	135·4	133·5	136·2	135·3
	Post treatment	136·3	134·3	135·9	135·2
lasma chloride concentration (mmol/l)	Pre-treatment	102·1	100·7	101·8	101·4
	Journey 1	101·6	100·5	102·2	101·9
	Lairage	102·4	101·3	102·4	102·0
	Journey 2	101·6	100·1	101·7	101·4
	Post treatment	102·3	101·1	102·4	101·4
ive weight (kg)	Pre-treatment	48·1	50·1	49·8	49·2
	Day 20 post treatment	55·0	56·2	55·2	55·0
o. of calves		22	24	24	24

' Different superscripts within a row indicate significant differences between treatments within an experimental period ' < 0·05).

' Different superscripts within a column indicate significant differences between journey 1 and journey 2 within a eatment ($P < 0·05$).

asma creatine kinase activity

eatment comparisons during experimental periods ıowed that the increase in the mean plasma creatine nase activity from pre-treatment values was gnificantly greater in transported calves than in fed ntrols during the lairage period, during journey 2 ıd during the 24-h post-treatment period (Figure 2; < 0·05).

ehydration indicators

ıcked cell volume. There were significant effects of eatment and period ($P < 0·001$) but no significant eatment × period interactions.

asma albumin concentration. There were no gnificant lairage duration or treatment effects during the experimental periods. However, after 8·75 h of journey 1, the change in plasma albumin concentration from pre-treatment values was significantly greater among transported calves (to 32·6 g/l) than among non-transported calves (to 31·2 g/l) ($P < 0·05$).

Plasma osmolality. Although there was a significant lairage × treatment × period interaction ($P < 0·05$), there were no significant treatment effects during the experimental periods.

Plasma sodium concentration. Although there was a significant treatment × period interaction ($P < 0·05$), there were no significant treatment differences during the experimental periods. However, during

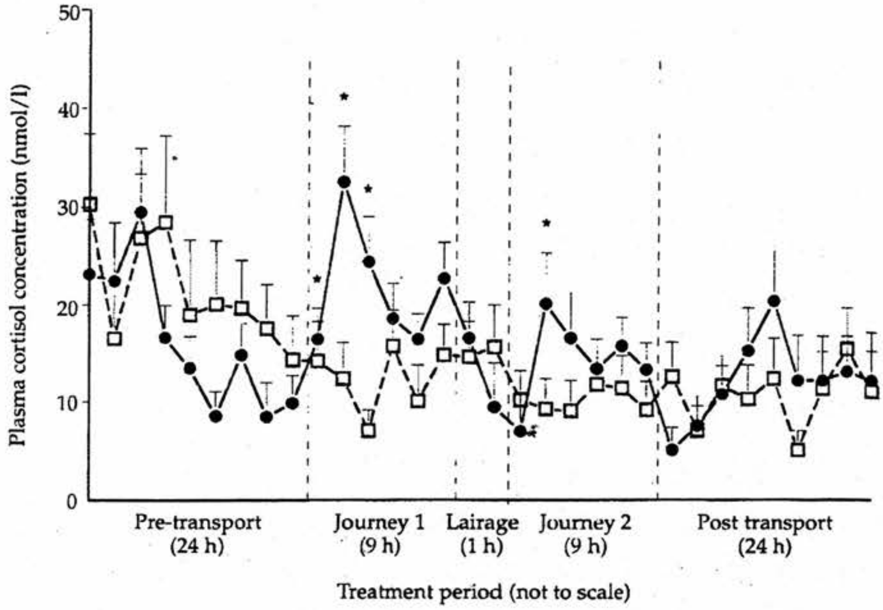


Figure 1 Effect of transport with a 1-h mid-journey lairage period on the mean plasma cortisol concentration: transported calves (—●—) and non-transported controls (—□—). Vertical bars show s.e.

journey 1, the plasma sodium concentration was significantly greater after 5·25 h of transport in transported calves (136 mmol/l) than in unfed controls (133 mmol/l) ($P < 0·05$), and after 8·75 h of transport it was significantly greater in transported calves (137 mmol/l) than in fed controls (135 mmol/l) ($P < 0·01$).

Plasma potassium concentration. Although there was a significant lairage × treatment × period interaction

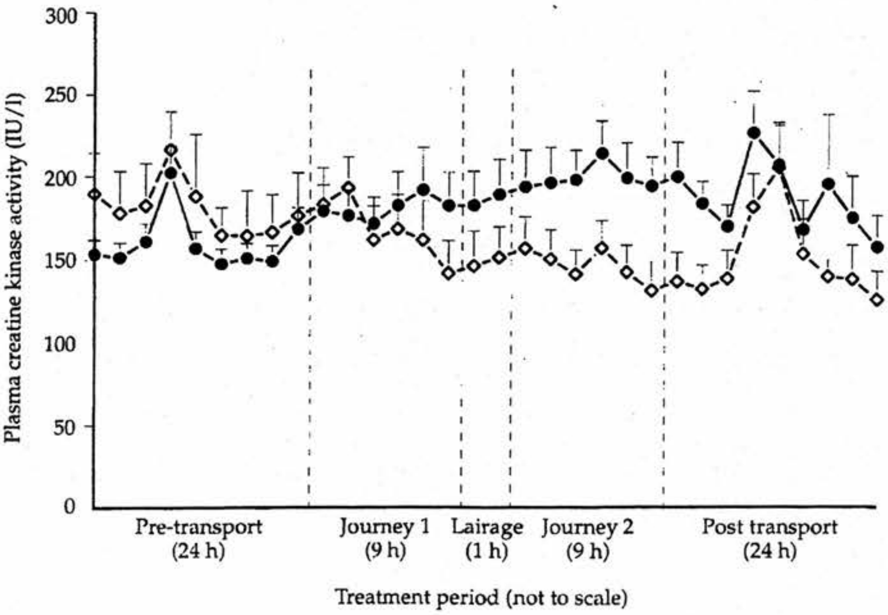


Figure 2 Effect of transport with a 1-h mid-journey lairage period on the mean plasma creatine kinase activity: transported calves (—●—) and non-transported fed controls (—◇—).

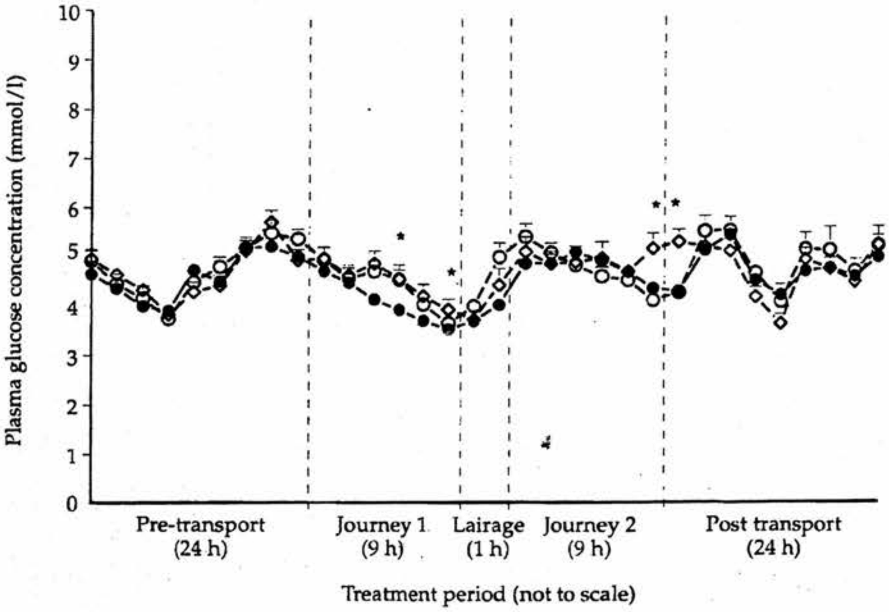


Figure 3 Effect of transport and food restriction with a 1-h mid-journey lairage period on the mean plasma glucose concentration: transported calves (—●—) and non-transported unfed controls (—○—) and non-transported fed controls (—◇—).

($P < 0{\cdot}001$), there were no significant treatment differences during the experimental periods.

Plasma chloride concentration. There were no significant treatment effects during the experimental periods.

Nutritional indicators

Plasma glucose concentration. During journey 1, there was a greater decrease in the mean plasma glucose concentration from pre-treatment values in transported than in fed control calves ($P < 0{\cdot}05$; Figure 3).

Plasma free fatty acid concentration. During both journey 1 and journey 2, the difference between the pre-treatment free fatty acid concentration and the mean concentration during transport was greater in transported calves than in either fed or unfed controls ($P < 0{\cdot}01$; Figure 4). After 8·75 h of food restriction, the plasma free fatty acid concentration in unfed controls was significantly greater than in fed controls ($P < 0{\cdot}05$).

Lying behaviour

During transport, the calves spent proportionately 0·56 of observations standing stationary, proportionately 0·44 of observations lying down and less than 0·01 of observations moving. As the time spent standing and time spent lying down were inversely proportional, only the data for time spent lying down were statistically analysed. Treatment comparisons during experimental periods showed that, during both journey 1 and journey 2, transported calves spent significantly less time lying down than control calves ($P < 0{\cdot}01$; Figures 5 and 6). There were no significant effects of space allowance during transport on the time spent lying down during the transport period. However, during the fourth circuit of journey 1, calves transported at the 0·475 m^2 per calf space allowance spent significantly more time lying down than those transported at the 0·375 m^2 per calf space allowance ($P < 0{\cdot}01$; Figure 7). During the lairage period, transported calves spent significantly more time lying down than fed controls ($P < 0{\cdot}01$). In all treatments, calves spent a significantly greater proportion of observations lying down during the 12-h lairage period than during the 1-h lairage period ($P < 0{\cdot}01$; Figures 5 and 6). However, there was no effect of lairage duration on lying behaviour either during journey 2 or post treatment (Figures 5 and 6). Transported calves spent a greater proportion of time lying during journey 2 than in journey 1 ($P < 0{\cdot}001$). There were no overall treatment effects on the mean lying behaviour during the 24-h post-transport period. However, 6 to 9 h after journey 2, fed controls spent significantly less time lying down than either transported calves or unfed controls ($P < 0{\cdot}05$).

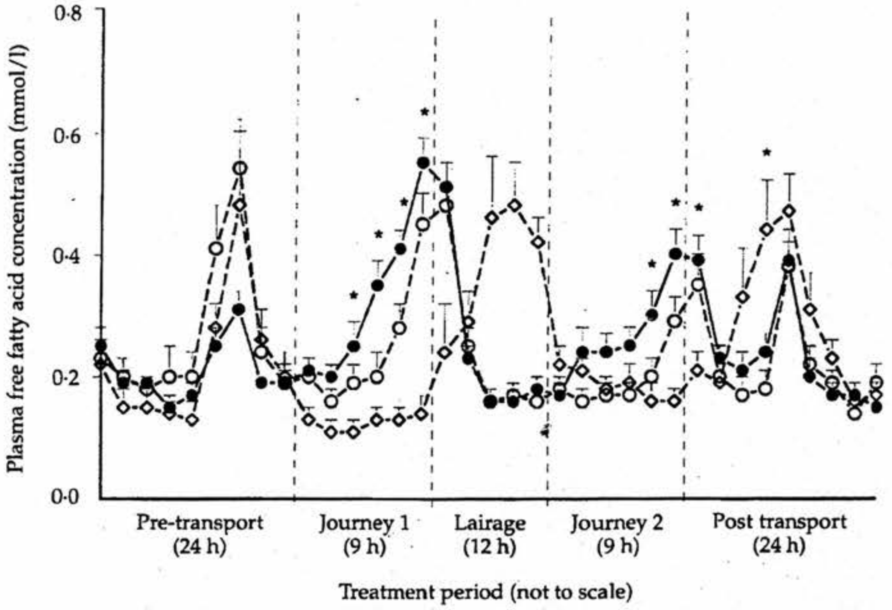


Figure 4 Effect of transport and food restriction with a 12-h mid-journey lairage period on the mean plasma free fatty acid concentration: transported calves (—●—) and non-transported unfed controls (—○—) and non-transported fed controls (—◇—).

Behaviour during transport and lairage
The calves engaged in few behavioural activities during transport and therefore treatment effects on these behaviours were not statistically analysed. The calves spent proportionately 0·13 of observations in an alert state i.e. with their head raised and ears

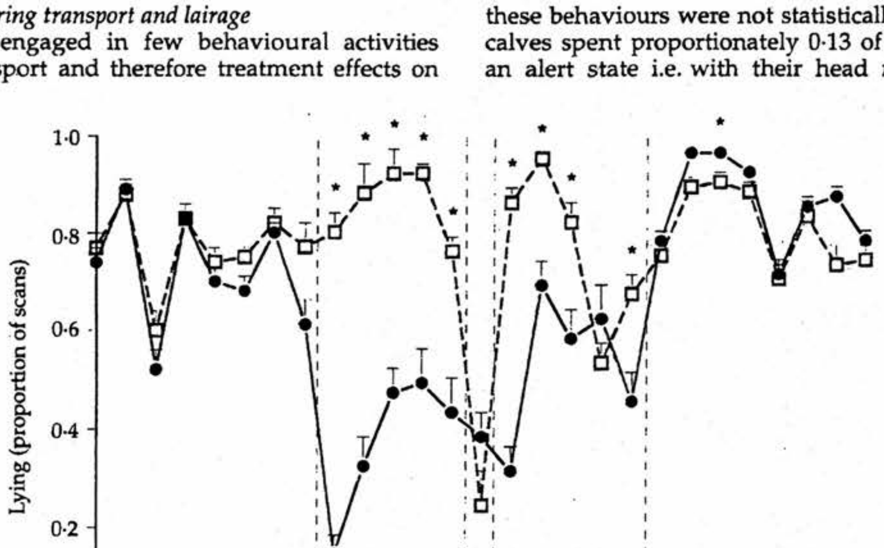
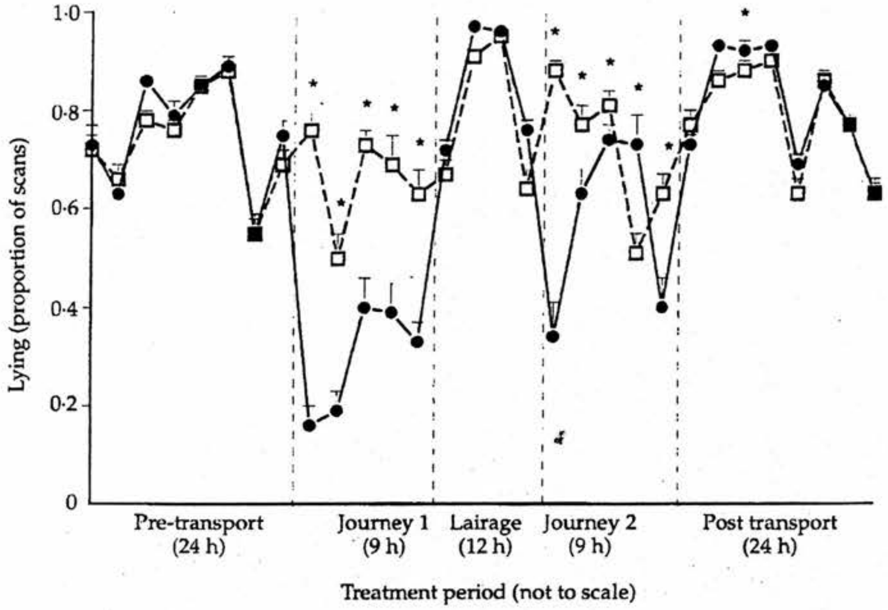
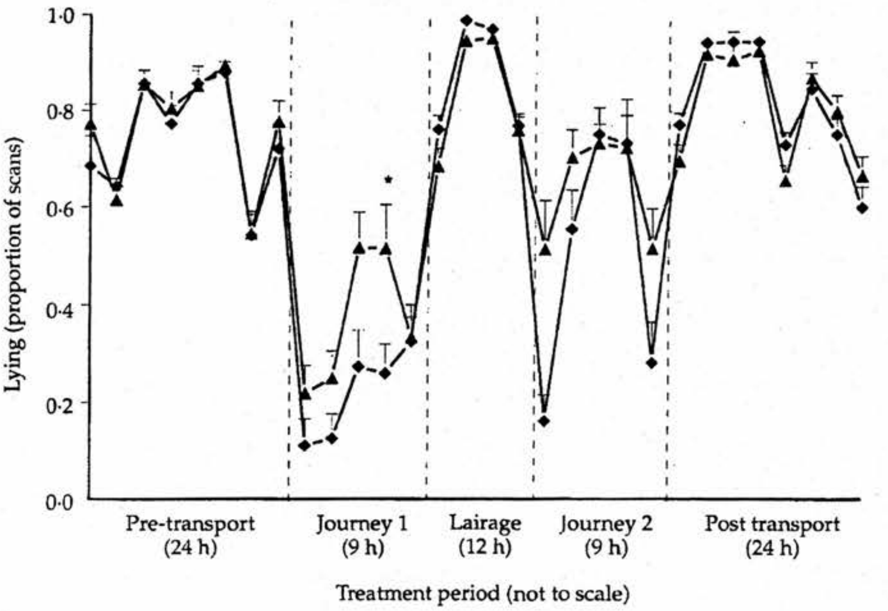


Figure 5 Effect of transport with a 1-h mid-journey lairage period on the mean proportion of scans per observation period (see text for details) during which the calves were observed lying down: transported calves (—●—) and non-transported controls (—□—). Vertical bars show s.e.



igure 6 Effect of transport with a 12-h mid-journey lairage period on the mean proportion of scans per observation period see text for details) during which the calves were observed lying down: transported calves (—●—) and non-transported ontrols (—□—). Vertical bars show s.e.

ɔricked, and 0·03 of observations performing oral nvestigation. They spent less than 0·01 of observations either ruminating, self-grooming or interacting with another calf. There was some



igure 7 Effect of space allowance during transport with a 12-h mid-journey lairage period on the mean proportion of scans er observation period (see text for details) during which the calves were observed lying down: transported calves space llowance of 0·375 m^2 per calf (—◆—) and at 0·475 m^2 per calf (—▲—).

evidence of an increase in alert behaviour (compared with pre-treatment values) among transported calves during journey 1 (mean proportion of observations was 0·17 compared with an overall mean of 0·09) but not during journey 2.

The most common orientation observed during transport was perpendicular to the direction of travel. At the 0·475 m^2 per calf space allowance, calves tended to orientate themselves either parallel to or perpendicular to the direction of travel (mean proportion of observations = 0·36 and 0·37, respectively). Calves transported at the 0·375 m^2 per calf space allowance tended to align themselves perpendicular to the direction of travel (mean proportion of observations = 0·54), whereas parallel orientations were less common (mean proportion of observations = 0·15).

During journey 1, there were no significant effects of space allowance on the median frequencies of losses of balance (4 events per calf per h), traumatic events i.e. falls + collisions with vehicle + collisions with another calf (1 event per calf per h), trampling (1 event per calf per h), and changes in posture (1 event per calf per h). During journey 2, when space allowance treatments were pooled and the effects of lairage duration were analysed at the batch level, there were significant effects of lairage duration. During journey 2, after a 1-h lairage, the median frequencies of losses of balance (4 events per calf per h) and trampling (2 events per calf per h) were greater than after the 12-h lairage (median frequencies of losses of balance were 2 events per calf per h and trampling were 0·5 events per calf per h) ($P < 0·05$).

Latency to drink water, drink milk replacer and lie down post transport

The mean time taken by transported calves to first drink (put their head in the water bucket) when they were returned to their pens after either journey was 6 min (range 0 to 54 min). Proportionately 0·57 of the transported calves drank water before lying down and there was no apparent effect of lairage duration on the proportion of calves that drank water before lying down. When offered milk replacer after a journey, the time taken by most transported calves to start to drink after the bucket was placed in front of the pen was between 0 and 35 s. However, after journey 1, seven transported calves were assisted to drink milk replacer (because they did not get up from a lying posture when the milk replacer was offered), compared with one unfed control. After journey 2, six transported calves were assisted to drink milk replacer, compared with two unfed controls. There was no significant difference between journeys 1 and 2 in the mean (23 min, range 0 to 93 min) time taken to lie down when the calves were returned to their pens. There was no effect of lairage duration on the time taken to lie down after journey 2. However the time taken for the calves to first lie down in their pens after journey 2 was shorter in calves that had been transported at the 0·375 m^2 per calf space allowance (mean latency was 966 s) than in those that had been transported at the 0·475 m^2 per calf space allowance (mean latency was 2022 s) ($P < 0·01$).

Heart rate

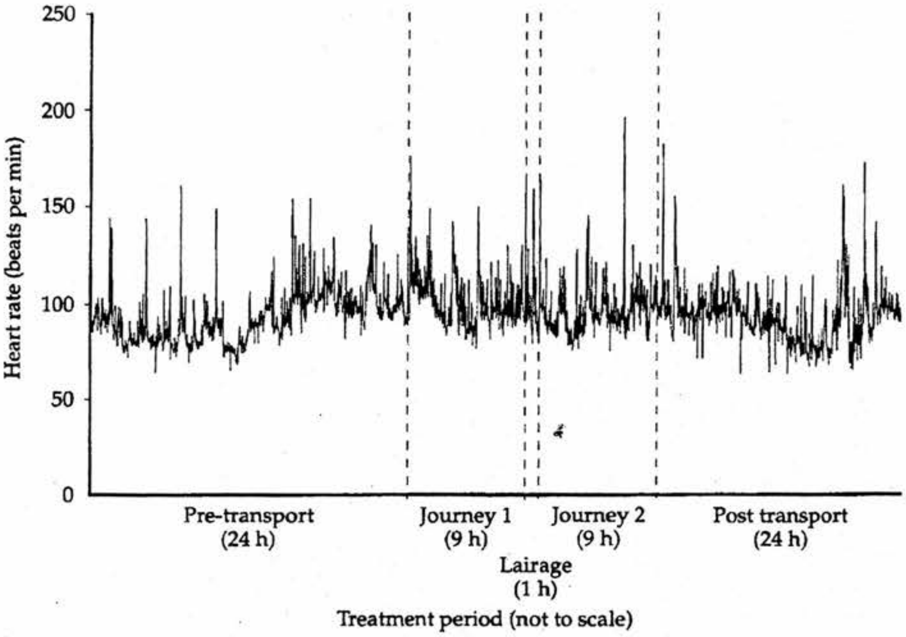
Although there were significant differences between some treatments in the change in heart rate from pre-treatment values during journey 1, these differences were not consistent and there was no difference between transported calves (mean 82 beats per min) and fed control calves (mean 86 beats per min). The heart rate values and pattern were similar during the transport periods to those pre- and post-transport, as shown by the example in Figure 8.

Live weight, food and water intakes

Transported calves lost proportionately 0·03 of their live weight during the transport and lairage treatment period, whereas fed controls gained proportionately 0·02 of their live weight during this period ($P < 0·01$). There were no other significant treatment effects on live-weight change during either the transport and lairage period or the post-transport period. With very few exceptions, the calves consumed all the milk replacer that they were offered and there were no significant treatment effects on concentrate intake during the post-treatment period (mean intake was 255 g per calf per day). There was no significant difference between transported calves and fed controls in the change in water intake from pre-treatment to that either during lairage or during the first 24-h post treatment. There was a decrease in water intake (by 0·08 l per calf per meal) from the pre-treatment values during the first 24-h post treatment in transported calves but not in unfed controls ($P < 0·05$).

Immunological measurements

BHV-1 infectivity in nasal swabs. BHV-1 was not detected in the nasal swabs from all of the calves. Only two calves from the fourth batch had detectable BHV-1 in the nasal swabs, so this batch was excluded from the statistical analysis. The proportion of calves with detectable BHV-1 in the nasal swabs was significantly greater in transported calves (0·81) than in control calves (0·58) (chi-square = 3·93, d.f. = 1, $P < 0·05$). Among calves with detectable BHV-1 in the nasal swabs, the concentration increased between 1 and 5 days after the start of the transport period and



igure 8 Example of heart rate recording from a calf transported at 0·475 m^2 per calf space allowance with a 1-h mid-journey irage period.

ıere was a significant treatment × period interaction $P < 0{\cdot}05$). However, there were no significant ifferences between transported calves and controls.

ymphocyte stimulation assay. There was no ymphocyte stimulation response to the BHV-1 ntigen and no treatment effects on the mean ymphocyte response to Con A at either 6 or 13 days ost transport.

erum BHV-1 antibody titre. Although there was a irage × treatment × period interaction ($P < 0{\cdot}01$), here were no significant treatment effects on the hange in the mean serum BHV-1 antibody titre from re-treatment to that at either 4, 9, 14 or 19 days post ransport.

erum ovalbumin antibody titre. Although there was a lear antibody response to the injection of valbumin, there were no significant treatment ffects on the change in the mean serum antibody itre from pre-treatment to that at either 4, 9, 14 or 19 lays post transport.

Iealth

ectal temperature. Although rectal temperatures aken during lairage were similar to those pre-reatment, during the 1st week post transport there vas an increase in rectal temperature from the pre-treatment values in transported calves but not in control calves ($P < 0{\cdot}05$; Figure 9). There were no significant treatment effects on rectal temperature during the subsequent post-treatment period.

Respiratory disease. There was no effect of either lairage duration or space allowance during transport on the incidence of respiratory disease (either number of calves with clinical signs of respiratory disease or number of calves treated for respiratory disease) in transported calves during the post-transport period. There was no significant effect of transport on the proportion of calves that received veterinary treatment for respiratory disease during the post-transport period (percentage of transported calves that received veterinary treatment for respiratory disease was 48% and the percentage of control calves that received veterinary treatment was 30%). However, as shown by Figure 10, there was some evidence that transport increased the risk of respiratory disease. When respiratory disease was classified as the occurrence of at least one of the following clinical signs: increased respiration rate, obvious serous or mucous nasal discharge, obvious serous or mucous ocular discharge or cough, the relative risk that transport was associated with respiratory disease during the 2- to 8-day post-transport period was 1·335 ($P_{0.95}$ confidence intervals were 1·020 to 1·746) and during the 9- to 20-day post-

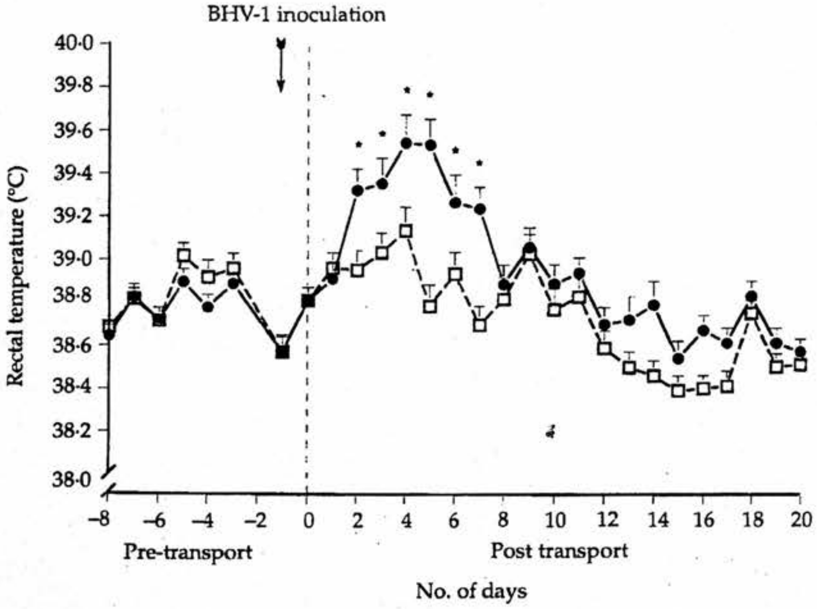


Figure 9 Effect of transport on the mean rectal temperature of calves: transported calves (—●—) and non-transported controls (—□—). Vertical bars show s.e. Vertical arrow indicates the timing of the BHV-1 intranasal inoculation.

transport period was 1·221 ($P_{0.95}$ confidence intervals were 1·001 to 1·490). Although there was a tendency for more transported calves (69%) to show gross pneumonic changes in the lungs *post mortem* than

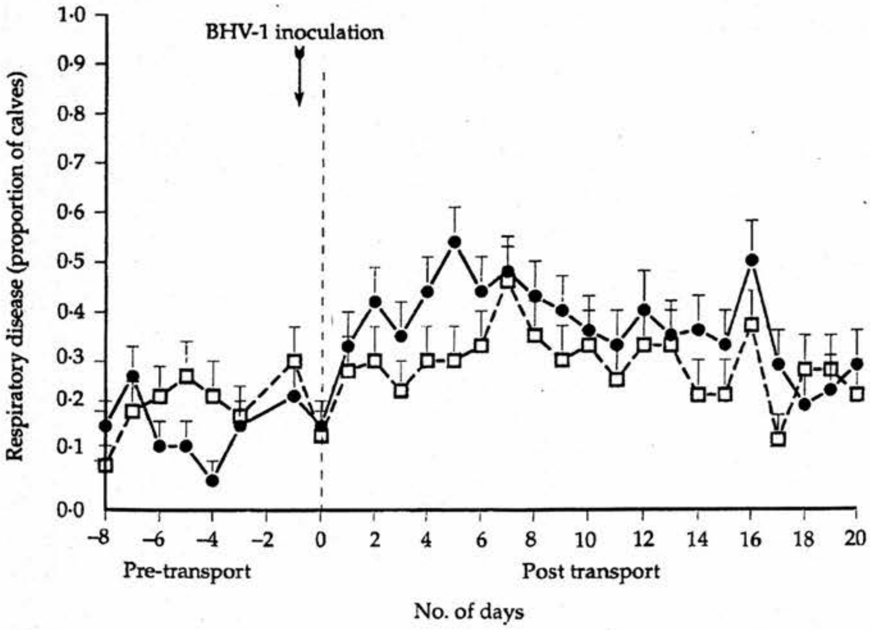


Figure 10 Effect of transport on the mean proportion of calves with clinical signs of respiratory disease: transported calves (—●—) and non-transported controls (—□—). Vertical bars show s.e. Vertical arrow indicates the timing of the BHV-1 intranasal inoculation.

ontrol calves (55%) this difference was not :atistically significant. The rectal temperature of ansported calves during the 1st week post-ansport was significantly correlated with the 1st 'eek post-transport clinical respiratory disease score '$_s$ 0·649, $P < 0·01$). However there were no other gnificant correlations between: post-transport ieasures of immunity and clinical signs of disease; ost-transport measures of immunity and stress ›sponses to transport; and post-transport clinical .gns of respiratory disease and stress responses to 'ansport.

)ehydration and diarrhoea. There was no effect of ither lairage duration or space allowance during :ansport on the incidence of either dehydration :linical signs of dehydration i.e. sunken eyes and ›ss of skin elasticity) or diarrhoea (faecal onsistency) in transported calves during the post-'ansport period. There was also no effect of :ansport on the incidence of either dehydration or iarrhoea post-transport (percentages of transported nd control calves with respectively either signs of .ehydration or liquid faeces were 42% and 46% .uring the 1st week post transport and 12% and 11% :om 9 to 20 days post transport). Only three calves ›quired veterinary treatment for dehydration .uring the post-treatment period.

)iscussion

'he plasma cortisol response to the first 9-h road ›urney indicated that one or more components of ne transport treatment were a stressor for young alves. The absence of a transport effect on the mean ieart rate was consistent with previous studies 3tephens and Toner, 1975; Knowles *et al.*, 1997). 'here was no evidence that food and water vithdrawal for a 9-h period affected either the ›lasma cortisol concentration or the heart rate. Cinsbergen *et al.* (1994) also found no effect of fasting alves for up to 24 h on either plasma cortisol oncentration or heart rate. As in studies of imulated transport (Agnes *et al.*, 1990), the peak in he plasma cortisol concentration occurred during he first part of the journey, rather than at the end of he journey. Although a delayed response to either ›ading, change from single to group penning or onfinement may have contributed to the peak ortisol response measured after the first circuit, here was no evidence of a treatment effect in the ›lood sample taken after loading but before the start ›f the journey. The decrease in plasma cortisol :oncentration between the two journeys suggests :ither that the second 9-h journey was a less stressful :xperience than the first journey or that the calves vere less able to produce an adrenal response during he second journey.

There was some evidence from raised plasma creatine kinase activity during transport to indicate that the calves experienced either increased muscular exertion or trauma associated with transport. The lack of a significant difference in the plasma creatine kinase activity between transported calves and controls during journey 1, but a significant difference between these treatments during journey 2 and the post-transport period, may have been due to the relatively slow release and slow clearance rate of the enzyme (Anderson *et al.*, 1976).

The behaviour of the calves immediately after transport was variable and suggested that some of the calves were either thirsty and seeking drinking water, or hungry and seeking milk replacer; however, for some calves the priority was to lie down and rest. In the environmental conditions of this experiment, the calves did not show any biochemical or clinical signs of dehydration either during the two 9-h journeys or during the two 9-h periods without food and water. The significant differences between transported and control calves in the plasma concentrations of albumin and sodium in the last blood sample taken during journey 1 occurred because of slight falls in the plasma concentrations of the control calves rather than because of an increase in the transported calves. The plasma concentrations of albumin and sodium, and the plasma osmolality were within their normal ranges and not indicative of dehydration (Kaneko *et al.*, 1997). In transported calves, the lower post-transport water intake compared with that pre-transport, suggests that the calves did not drink additional water to replace a water deficit accumulated during transport.

The major factor that appeared to affect the plasma free fatty acid concentration was the time since the last meal. Peak plasma free fatty acid concentrations coincided with the longest time since feeding. This finding is consistent with Kinsbergen *et al.* (1994) who also showed an increase in the plasma free fatty acid concentration during fasting and a subsequent fall after feeding. In fasted calves, a rise in plasma free fatty acids is associated with a mobilization of fat reserves in response to an energy deficiency (Kinsbergen *et al.*, 1994; Felber and Golay, 1995; Penicaud *et al.*, 2000). A second, but less important factor affecting the plasma free fatty acid concentration was the transport treatment. In transported calves, the plasma free fatty acid concentration rose earlier and reached a higher peak during both journey 1 and journey 2 than in unfed controls. This may have been due to a stress response to the transport treatment resulting in increased plasma catecholamine concentrations that mobilized fat reserves (Frohli and Blum, 1988). However, the greatest difference in the plasma free fatty acid concentration between transported and unfed controls was at the end of the journey, whereas the stress response was greatest at the start of the journey, suggesting that this effect was more likely to have been due to a greater energy demand in the transported calves compared with that in the non-transported unfed controls. During journey 1 there was a fall in the plasma glucose concentration, however the rectal temperatures taken during lairage were similar to those pre-treatment indicating that the calves had not become hypothermic as a result of an energy deficiency during transport. Differences in feeding times and the rise in the plasma glucose concentration that follows after feeding (Kinsbergen *et al.*, 1994), are the likely explanations for the differences in the plasma glucose concentration both between treatment groups and between journey 1 and journey 2.

Although at both space allowances there was sufficient floor space for the calves to lie down (in both space allowances, for short periods of time all calves in a pen lay down), there was less lying behaviour in transported calves than in control calves. This suggests that some aspect of transportation was inhibiting lying behaviour. As the duration of transport increased, either within a journey or during the second journey, there was more lying behaviour which suggests that either the calves became habituated to some extent or became fatigued and the motivation to lie down overcame the inhibitory influences of transportation. During the lairage period, transported calves spent significantly more time lying down than fed controls and the amount of lying behaviour was greater during journey 2 than during journey 1. This suggests that the calves may have become fatigued; however during journey 2, the mean proportion of observations during which the calves were lying down was less than 0·6 and although 3 to 12 h post transport it was greater than 0·9, there was no significant difference between the post-transport lying behaviour of transported and control calves. The mean time of 23 min (range 0 to 93 min) taken by transported calves to first lie down when they were returned to their pens after either journey 1 or journey 2 suggests that most calves were not completely exhausted after their journey. Transported calves tended to take longer to lie down after journey 2 than after journey 1, possibly due to the increased time spent lying down during journey 2 compared with journey 1. Fit and healthy calves would normally drink their milk replacer almost immediately. The delay of some of the transported calves in getting up from a lying posture to drink and the failure of some of these calves to voluntarily get up from a lying posture to drink is an indicator that at least some of the transported calves were fatigued after the journey.

The most common orientation observed during transport (perpendicular to the direction of travel) may have afforded the most stability during transport. In both journeys, calves transported at the lower space allowance tended to align themselves less often either directly towards or away from the direction of travel than calves at the higher space allowance. This may have been a consequence of reduced floor area to adjust their posture, but it may also have been a consequence of the lower pen length at the lower space allowance than at the higher space allowance. At the greater space allowance, which allowed calves more opportunity to adopt their preferred orientations, calves tended to orientate themselves either perpendicular or parallel to the direction of travel.

There were few significant effects of space allowance on the behavioural and physiological responses of the calves to transport. There was no evidence that the calves were at greater risk of injury at the higher space allowance due to loss of stability in response to vehicle movements.

There were few significant effects of lairage duration on the physiological and behavioural responses of the calves. To compare the effects of lairage duration at equivalent times of day during journey 2 and post transport, journey 1 had to be started at 12:00 h for the 12-h lairage treatment and at 23:00 h for the 1 h lairage treatment. These differences in starting time for journey 1 could have confounded any effects of lairage duration, however, there was no evidence that this was the case. Although the median number of starts/stops during journey 1 was five with a starting time of 12:00 h and two with a starting time of 23:00 h, there were no significant differences between the starting times in the median frequency of losses of balance or traumatic events experienced by the calves during journey 1. It was also possible that the responses to journey 1 could have been influenced by circadian rhythms; however, the patterns of both lying behaviour and plasma cortisol concentration were similar for each starting time. A 12-h mid-journey lairage period provided additional time for calves to lie down and rest as compared with a short mid-journey lairage period of 1 h. However, a 1-h mid-journey lairage period compared with a 12-h mid-journey lairage did not appear to be detrimental to the welfare of the calves during the subsequent 9-h journey or post transport. The only evidence that the calves were more fatigued after the norter lairage period was a slight increase in the requency of losses of balance and trampling during he second journey. Although the calves did spend ignificantly more time lying down during the 12-h airage period than during the 1-h lairage period, here was no effect of lairage duration on lying ehaviour either during the subsequent 9-h journey r post treatment.

As there were no significant effects of a 9-h period vithout food and water on the biochemical neasurements of dehydration, there was no major enefit in providing a prolonged time to drink luring the lairage period. There was no significant lifference in the free fatty acid response of the calves o a subsequent 9-h transport period after they were ed once during a 1-h mid-journey lairage period compared with being fed both at the start and the nd of a 12-h mid-journey lairage period. Offering calves milk replacer, either immediately before or ifter transport did not result in diarrhoea and lehydration. However, in the circumstances of this experiment, the milk replacer offered during lairage and immediately post transport was the same nutritional composition, was prepared in the same way and was offered at the same feeding level, as before transport.

Another difference between the lairage treatment here and that likely to be experienced by calves transported commercially was that for experimental reasons, the transported calves were returned to their original single pens and were individually offered food and water by bucket, whereas commercial practice would normally involve group penning in a novel environment and food and water would be offered by either bucket, trough or teat. If group penned, the behaviour of the calves is likely to be influenced by social facilitation and social interactions. However, if legislation for staging points (European Council, 1997) is followed, there should be facilities and staff to ensure that each liquid fed calf is offered water and sufficient appropriate food to satisfy its bodily needs during its stay and for the expected duration of its journey to the next feeding point.

Although there was an obvious cortisol response to the transport treatment, there was little evidence of immunosuppression post transport. The absence of both a lymphocyte stimulation response to the BHV-1 antigen and a detectable antibody response to the BHV-1 virus inoculation, but virus recovery from nasal swabs and a clinical response after BHV-1 inoculation was most likely due to the presence of maternal antibodies in some of the calves (Bradshaw and Edwards, 1996). It was expected that the use of the Oxford strain of BHV-1 would produce an immunological response in calves with maternal antibodies and permit virus recovery, particularly if the calves had become immunosuppressed as a result of the transport treatments. Although in the circumstances of this study, the immunological responses of the calves to inoculation with BHV-1 did not provide a suitable model on which to evaluate the effects of different transport treatments, the other immunological responses that were not affected by the presence of maternal antibodies did allow the treatment effects to be assessed. However, there were no treatment effects on the general responsiveness of either the humoral or the cellular components of the immune system.

There was some evidence that the transport treatment adversely affected the health of the calves post transport. The effects of the transport treatment were raised rectal temperatures and greater clinical signs of respiratory disease post transport. The rise in rectal temperatures in the transported calves compared with pre-treatment levels and with non-transported controls indicates a greater clinical response to either the BHV-1 inoculation or other infections. There was no evidence that the inoculation resulted in either pyrexia or other clinical signs during transportation and it was unlikely that the BHV-1 inoculation affected the plasma cortisol concentration of the calves (Arthington *et al.*, 1997). However, the post-transport clinical response of the calves to the BHV-1 inoculation may have exaggerated some effects of the transport treatment (e.g. calves with an active infection after BHV-1 inoculation can experience a negative nitrogen balance (Orr *et al.*, 1988)) and the risk of post-transport respiratory disease in the transported calves may have been greater than in healthy calves that had not been exposed to a viral challenge before transportation. Respiratory disease can reduce live-weight gain (Donovan *et al.*, 1998) and the respiratory disease that occurred during the post-transport period may have affected the subsequent live-weight gain (0·3 kg/day). Although the calves lost live-weight during the transport and lairage treatments and there was some evidence that the transport treatment affected health, there was no effect on subsequent live-weight gain.

The main conclusions from this study are that an aspect of the transport treatment was a stressor to the calves, and the effect of this stressor appeared to decrease with journey duration and during a subsequent journey. With the type of feeding used, the time since the last meal was an important factor in the response of the calves to the journey, as during a 9-h journey the calves mobilized free fatty acids, probably in response to an energy deficiency. Under the environmental conditions of this experiment, there was no evidence that lack of drinking water was an important factor during a 9-h journey. Under the driving conditions used, and the range of space allowances studied (0·375 to 0·475 m^2 per calf), increasing the space allowance was not associated with a greater loss of stability and greater risk of injury. The duration of the mid-journey lairage was not an important factor, as although a longer lairage period did provide greater opportunity for rest, a short lairage duration of 1 h, sufficient for the calves to receive milk replacer, but with little opportunity for the calves to rest, had no major detrimental effects on the variables used to assess the welfare of the calves. Although there was no evidence that transport affected the immunological variables measured during the post-transport period, there was some evidence that transport adversely affected the health of the calves post transport.

Acknowledgements

This work has been financed by the European Commission, Directorate General for Agriculture, within the framework of the RTD contract FAIR 2049: 'Chain Management of Veal Calf Welfare'. The content of this publication is the sole responsibility of its publishers. It in no way represents the views of the Commission or its services. The Department of Veterinary Clinical Studies Diagnostic Laboratory performed some of the biochemical analyses. British Denkavit assisted with the cost of the food. The Moredun Research Foundation provided the accommodation and care for the calves.

References

Agnes, F., Sartorelli, P., Abdi, B. H. and Locatelli, A. 1990. Effect of transport loading or noise on blood biochemical variables in calves. *American Journal of Veterinary Research* **51:** 1679-1681.

Anderson, P. H., Berret, S. and Patterson, D. S. P. 1976. The significance of elevated plasma creatine phosphokinase activity in muscle disease of cattle. *Journal of Comparative Pathology* **86:** 531-538.

Arthington, J. D., Corah, L. R., Minton, J. E., Elsasser, T. H. and Blecha, F. 1997. Supplemental dietary chromium does not influence ACTH, cortisol, or immune responses in young calves inoculated with bovine herpesvirus-1. *Journal of Animal Science* **75:** 217-223.

Atkinson, P. J. 1992. Investigation of the effects of transport and lairage on hydration state and resting behaviour of calves for export. *Veterinary Record* **130:** 413-416.

Blecha, F., Boyles, S. L. and Riley, J. G. 1984. Shipping suppresses lymphocyte blastogenic responses in Angus and Brahman x Angus feeder calves. *Journal of Animal Science* **59:** 576-583.

Bradshaw, B. J. F. and Edwards, S. 1996. Antibody isotype responses to experimental infection with bovine herpesvirus 1 in calves with colostrally derived antibody. *Veterinary Microbiology* **53:** 143-151.

Burrells, C., Inglis, N. F., Davies, R. C. and Sharp, J. M. 1995. Detection of specific T-cell reactivity in sheep infected with *Mycobacterium avium* subspecies *silvaticum* and paratuberculosis using two defined bacterial antigens. *Veterinary Immunology and Immunopathology* **45:** 311-320.

Burrells, C. and Wells, P. W. 1977. *In-vitro* stimulation of ovine lymphocytes by various mitogens. *Research in Veterinary Science* **23:** 84-87.

Cockram, M. S., Kent, J. E., Goddard, P. J., Waran, N. K., McGilp, I. M., Jackson, R. E., Muwanga, G. M. and Prytherch, S. 1996. Effect of space allowance during transport on the behavioural and physiological responses of lambs during and after transport. *Animal Science* **62:** 461-477.

Cockram, M. S. and Mitchell, M. A. 1999. Role of research in the formulation of rules to protect the welfare of farm animals during road transportation. In *Farm animal welfare-who writes the rules ?* (ed. A. J. F. Russel, C. A. Morgan, C. J. Savory, M. C. Appleby and T. L. J. Lawrence), *British Society of Animal Science occasional publication no. 23*, pp. 43 -64.

Constable, P. D., Walker, P. G., Morin, D. E. and Foreman, J. H. 1998. Clinical and laboratory assessment of hydration status of neonatal calves with diarrhea. *Journal of the American Veterinary Medical Association* **212:** 991-996.

Donovan, G. A., Dohoo, I. R., Montgomery, D. M. and Bennett, F. L. 1998. Calf and disease factors affecting growth in female Holstein calves in Florida, USA. *Preventive Veterinary Medicine* **33:** 1-10.

European Council. 1991. Council directive 91/628/EEC of November 19th 1991 on the protection of animals during transport and amending directives 90/425/EEC and 91/496/EEC. *Official Journal of the European Communities* **L340:** 17-38.

European Council. 1995. Council directive 95/29/EC of June 1995 amending directive 91/628/EC concerning the protection of animals during transport. *Official Journal of the European Communities* **L148:** 52-63.

European Council. 1997. Council regulation (EC) no. 1255/97 of 25 June 1997 concerning Community criteria for staging points and amending the route plan referred to in the annex to directive 91/628/EEC. *Official Journal of the European Communities* **L174:** 1–6.

Felber, J. P. and Golay, A. 1995. Regulation of nutrient metabolism and energy-expenditure. *Metabolism, Clinical and Experimental* **44:** 4–9.

Frohli, D. and Blum, J. W. 1988. Effects of fasting on blood plasma levels, metabolism and metabolic effects of epinephrine and norepinephrine in steers. *Acta Endocrinologica (Copenhagen)* **118:** 254-259.

Great Britain Parliament. 1997. *The Welfare of Animals (Transport) Order, 1997.* Statutory instrument 1997/1480. Her Majesty's Stationery Office, London.

Johnston, J. D. and Buckland, R. B. 1976. Response of male Holstein calves from seven sires to four management stresses as measured by plasma corticoid levels. *Canadian Journal of Animal Science* **56:** 727-732.

Kaneko, J. J., Harvey, J. W. and Bruss, M. L. 1997. *Clinical biochemistry of domestic animals.* Academic Press, San Diego, USA.

Kent, J. E. and Ewbank, R. 1986. The effect of road transportation on the blood-constituents and behaviour of calves. 2. One to 3 weeks old. *British Veterinary Journal* **142:** 131-140.

Kinsbergen, M., Sallmann, H. P. and Blum, J. W. 1994. Metabolic, endocrine and haematological-changes in 1-week-old calves after milk intake, in response to fasting and during total parenteral-nutrition. *Journal of Veterinary Medicine, Series A-Physiology, Pathology, Clinical Medicine* **41:** 268-282.

Knowles, T. G. 1995. A review of post transport mortality among younger calves. *Veterinary Record* **137:** 406-407.

Knowles, T. G. 1999. A review of the road transport of cattle. *Veterinary Record* **144:** 197-201.

Knowles, T. G., Brown, S. N., Edwards, J. E., Phillips, A. J. and Warriss, P. D. 1999. Effect on young calves of a one-hour feeding stop during a 19-hour road journey. *Veterinary Record* **144:** 687-692.

Knowles, T. G., Warriss, P. D., Brown, S. N., Edwards, J. E., Watkins, P. E. and Phillips, A. J. 1997. Effects of calves less than one month old of feeding or not feeding them during road transport of up to 24-hours. *Veterinary Record* **140:** 116-124.

Laird, N. M. and Ware, J. H. 1982. Random-effects models for longitudinal data. *Biometrics* **38:** 963-974.

Lyaku, J. R. S., Nettleton, P. F. and Scott, G. R. 1990. A quantitative enzyme-linked immunosorbent assay for bovine herpesvirus type 1 (BHV-1) antibody. *Biologicals* **18:** 199-205.

Mormede, P., Soissons, J., Bluthe, R. M., Raoult, J., Legarff, G., Levieux, D. and Dantzer, R. 1982. Effect of transportation on blood-serum composition, disease incidence, and production traits in young calves-influence of the journey duration. *Annales de Recherches Vétérinaires* **13:** 369-384.

Murata, H. 1989. Suppression of lymphocyte blastogenesis by sera from calves transported by road. *British Veterinary Journal* **145:** 257-262.

Murata, H. and Hirose, H. 1990. Impairment of lymphocyte blastogenesis in road-transported calves observed with a whole blood culture technique. *Japanese Journal of Veterinary Science* **52:** 183-185.

Murata, H. and Hirose, H. 1991. Suppression of bovine lymphocyte and macrophage functions by sera from road-transported calves. *British Veterinary Journal* **147:** 455-462.

Noldus Information Technology. 1993. *The Observer, base package for DOS. Reference manual, version 3·0 edition*. Noldus Information Technology, Wageningen, The Netherlands.

Noldus Information Technology. 1994. *The Observer, support package for the Psion Organiser. User's manual, version 3·0.* Noldus Information Technology, Wageningen, The Netherlands.

Orr, C., Hutcheson, D. P., Cummins, J. M. and Thompson, G. B. 1988. Nitrogen kinetics of infectious bovine rhinotracheitis-stressed calves. *Journal of Animal Science* **66:** 1982-1989.

Penicaud, L., Cousin, B., Leloup, C., Lorsignol, A. and Casteilla, L. 2000. The autonomic nervous system, adipose tissue plasticity, and energy balance. *Nutrition* **16:** 903-908.

Roth, J. A. and Kaeberle, M. L. 1982. Effect of glucocorticoids on the bovine immune system. *Journal of the American Veterinary Medical Association* **180:** 894-901.

Roth, J. A., Kaeberle, M. L. and Hsu, W. H. 1982. Effects of ACTH administration on bovine polymorphonuclear leukocyte function and lymphocyte blastogenesis. *American Journal of Veterinary Research* **443:** 412-416.

Staples, G. E. and Haugse, C. N. 1974. Losses in young calves after transportation. *British Veterinary Journal* **130:** 374-379.

Stephens, D. B. and Toner, J. N. 1975. Husbandry influences on some physiological parameters of emotional responses in calves. *Applied Animal Ethology* **1:** 233-243.

Tarrant, P. V. 1990. Transportation of cattle by road. *Applied Animal Behaviour Science* **28:** 153-170.

Trunkfield, H. R. and Broom, D. M. 1990. The welfare of calves during handling and transport. *Applied Animal Behaviour Science* **28:** 135-152.

Wolfson, W. Q., Cohn, C., Calvary, E. and Ichiba, F. 1948. Studies in serum proteins. V. A rapid procedure for the estimation of total protein, true albumin, total globulin, alpha globulin, beta globulin and gamma globulin in 1·0 ml of serum. *American Journal of Clinical Pathology* **18:** 723-730.

Yates, W. D. G. 1982. A review of infectious bovine rhinotracheitis, shipping fever pneumonia and viral-bacterial synergism in respiratory disease of cattle. *Canadian Journal of Comparative Medicine* **46:** 225-263.

(*Received* 1 *September* 2000—*Accepted* 11 *April* 2001)

Articles

14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening

F E Alexander, T J Anderson, H K Brown, A P M Forrest, W Hepburn, A E Kirkpatrick, B B Muir, R J Prescott, A Smith

Summary

Background The Edinburgh randomised trial of breast-cancer screening recruited women aged 45–64 years from 1978 to 1981 (cohort 1), and those aged 45–49 years during 1982–85 (cohorts 2 and 3). Results based on 14 years of follow-up and 270 000 woman-years of observation are reported.

Methods Breast-cancer mortality rates in the intervention group (28 628 women offered screening) were compared with those in the control group (26 026) with adjustment for socioeconomic status (SES) of general medical practices. Rate ratios were derived by means of logistic regression for the total trial population and for women first offered screening while younger than 50 years. Analyses were by intention to treat.

Findings Initial unadjusted results showed a difference of just 13% in breast-cancer mortality rates between the intervention and control groups (156 deaths [5·18 per 10 000] *vs* 167 [6·04 per 10 000]; rate ratio 0·87 [95% CI 0·70–1·06]), but the results were influenced by differences in SES by trial group. After adjustment for SES, the rate ratio was 0·79 (95% CI 0·60–1·02). When deaths after diagnosis more than 3 years after the end of the study were censored the rate ratio became 0·71 (0·53–0·95). There was no evidence of heterogeneity by age at entry and no evidence that younger entrants had smaller or delayed benefit (rate ratio 0·70 [0·41–1·20]). No breast-cancer mortality benefit was observed for women whose breast cancers were diagnosed when they were younger than 50 years. Other-cause mortality rates did not differ by trial group when adjusted for SES.

Interpretation Our findings confirm results from randomised trials in Sweden and the USA that screening for breast cancer lowers breast-cancer mortality. Similar results are reported by the UK geographical comparison, UK Trial of Early Detection of Breast Cancer. The results for younger women suggest benefit from introduction of screening before 50 years of age.

Lancet 1999; **353:** 1903–08
See Commentary page 1896

Department of Community Health Sciences (F E Alexander PhD, H K Brown MSc, W Hepburn, R J Prescott PhD, A Smith), **Department of Pathology** (T J Anderson FRCPath), **and Hugh Robson Link Building** (Prof A P M Forrest MD), **University of Edinburgh, Edinburgh, UK; and South East Scotland Division, Scottish Breast Screening Programme, Edinburgh** (A E Kirkpatrick FRCR, B B Muir FRCR)

Correspondence to: Dr F E Alexander, Department of Community Health Sciences, University of Edinburgh, Edinburgh EH8 9AG, UK (e-mail: freda.alexander@ed.ac.uk)

Introduction

The Edinburgh randomised trial of breast-cancer screening was started in 1978.[1] Between 1978 and 1981, the trial recruited 44 288 women aged 45–64 years of age from 87 general practices in Edinburgh to form the initial cohort. Subsequently, further eligible women who became patients of these practices and existing patients who reached 45 years of age were recruited in two further cohorts, 4867 in 1982–83 (cohort 2) and 5499 in 1984–85 (cohort 3).[2] Participating practices were randomly assigned to intervention and control groups, women taking their status from their general medical practice at their time of entry. Women in the intervention group were invited to participate in a screening programme, which included clinical examination every year and two-view mammography every 2 years. Control-group women received only normal medical care. The prospectively defined hypothesis was that breast-cancer mortality would be lower in the intervention group than in the control group after 7 years and longer periods of follow-up. At that time, there was no reason to believe that the effect might differ by age of entry to the trial and no subgroup analyses were planned. Subsequent evidence from other randomised controlled trials of mammographic screening have placed importance on two age-groups (<50 years, ⩾50 years at entry).

There have been two reports[2,3] of mortality from breast-cancer relating to experience of women in the initial cohort followed up for 7 and 10 years, respectively. The second report also included the later cohorts over a shorter follow-up period. For both these follow-up periods, we report that breast-cancer mortality was 17–18% in the screening group; these differences were not significant. The 10-year analysis reported that breast-cancer mortality was 22% lower in the screening group for younger women aged 45–59 years at entry.

As a result of the cluster randomisation, there was bias between the two groups, women in the control group having higher all-cause mortality rates and lower socioeconomic status (SES) than those randomly assigned to intervention.[4] Attempts to adjust for these differences by quantifying SES in samples from each practice were unsuccessful.[2,3] An improved method of quantifying SES has now been developed, which we believe adjusts for this bias.[5] We include in this report the effect of this adjustment on breast-cancer mortality rates after 14 years and separately consider the effect of screening for younger women.

With longer periods of follow-up, the inclusion of women whose diagnosis could not have been influenced by screening becomes difficult. The consequences of different policies on censoring of deaths by their date of diagnosis were considered in the Swedish overview analysis[6] and the 14-year report of the Health Insurance Plan (HIP) trial[7] and are now considered for the first time for the Edinburgh randomised trial. The HIP investigators[7] were the first to note that for entrants younger than 50 years to the intervention group benefit was restricted to women whose breast cancers were diagnosed when they were in their fifties. Indeed, breast-cancer mortality for women whose breast cancers were diagnosed when they were younger than 50 years was higher in the intervention than in the control group. A corresponding analysis is now done for the Edinburgh trial. Women in the intervention group of the Edinburgh trial formed the Edinburgh component of the UK Trial of Early Detection of Breast Cancer (TEDBC).[8] The latest results of TEDBC are also reported in this issue.[9]

Methods

Study participants and design

The geographical base for this trial was 87 general medical practices within the city of Edinburgh.[1] The only practices not included were those that had participated in a pilot study or refused to take part. Every woman aged 45–64 years registered at one of these practices and without a previous diagnosis of breast cancer was eligible for entry to the trial. Practices were individually randomised after stratification by number of partners (to balance the numbers in the two groups) and entered sequentially between 1978 and 1981. All women aged 45–64 years from each practice were entered as the practice was recruited (with cluster randomisation of women) to form an initial cohort (cohort 1). During two 2-year periods (1982–83 and 1984–85), newly eligible women registered with participating practices were also recruited to form cohorts 2 and 3. We included only women aged 45–49 years at entry in these two cohorts. Older entrants to these cohorts were mainly women who had moved to Edinburgh and were not representative of the population. The only exclusion criterion applied at trial onset was a previous diagnosis of breast cancer, but this was normally ascertained after randomisation (figure 1). Informed consent was obtained for the screening process, but not for trial entry. The trial was approved by the Lothian Research Ethics Committee. The primary outcome measure was the rate ratio of breast-cancer mortality rates in the two groups; the trial was designed to have 80% statistical power to detect a 30% lower mortality rate in the intervention group after 7 years of follow-up (rate ratio ≤0·70), by a one-sided test.[1] Power calculations ignored the cluster randomisation.

The study continued until 1988, with women in the intervention group offered a maximum of four mammographic screenings every 2 years for cohort 1, three screenings for cohort 2, and two screenings for cohort 3. During the study period these women were also offered clinical breast examinations once a year. With the start of service screening in Edinburgh in June, 1988, all trial women younger than 65 years were eligible to be screened. Those who had participated in regular screening during the trial were invited for their first service screening 3 years after their last trial screening. Controls were invited for their first service screening at the corresponding time: year 10 from entry for the initial cohort, and years 8 and 6 for cohorts 2 and 3, respectively.

The records of all women in the trial were flagged by the General Registry Office in Edinburgh in 1985. This standard UK procedure allows follow-up information on cancer incidence, breast-cancer mortality, and other-cause mortality to be sent to the trial administrators wherever individual participants are living. This report is restricted to the 98% of women whose records were successfully flagged. Failure of flagging is therefore a second exclusion criterion; most of the women whose records were not flagged had died or left Edinburgh before the trial began, although their records remained on their general practice list.[1]

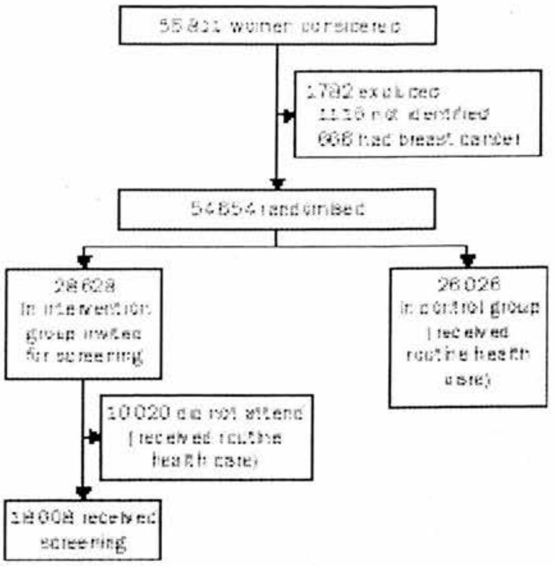

55821 women considered
1782 excluded
1115 not identified
668 had breast cancer
54654 randomised
28628 in intervention group invited for screening
26026 in control group (received routine health care)
10020 did not attend (received routine health care)
18608 received screening

Figure 1: **Trial profile**

Data analysis

Mortality statistics were based on death certification accessed through the flagging process. Breast-cancer deaths are those in which breast cancer was mentioned on the death certificate as the underlying cause, according to WHO rules.[10] All such deaths were checked against the trial database and, if necessary, additional sources. They were included only after confirmation that the date at which breast cancer was diagnosed was later than that of randomisation. All deaths in which breast cancer was not the underlying cause were recorded as other-cause mortality. The follow-up period is 14 years for cohort 1, 12 years for cohort 2, and 10 years for cohort 3, from date of randomisation. To counteract the effect of the inclusion of deaths from breast cancer that could not have been influenced by screening, results are presented not only for all deaths during follow-up, but also with death censored if diagnosis occurred after a specified date (1987, 1989, 1991) and if the diagnosis occurred after a specified number of years in the study. Additional analyses censored deaths after diagnoses before or after age 50 years to allow comparisons with results from the HIP trial.[7]

Breast-cancer and other-cause mortality rates were calculated as rates per 10 000 woman-years at risk for the two groups in the trial, and the risk ratio calculated with a modified logistic regression procedure, as in previous studies,[2,3] incorporating adjustment for extrabinomial variation by the method of Williams[11] to respect the cluster randomisation. These analyses were done in SAS Proc Logistic version 6.12 and stratified by age at randomisation (45–49 years, 50–54 years, 55–59 years, 60–64 years). When the later cohorts were included further stratification by cohort was added. Age-specific results were derived from fitting of interaction terms for age with trial group, but all women were included in the modelling process. Cumulative mortality curves were expressed as rates per 10 000 women entered, but were adjusted to take account of woman-years at risk. Our method of analysis differs somewhat from that used for TEDBC; that trial used Poisson rather than logistic regression, but we used the TEDBC method[9] to confirm that the differences in statistical methods do not explain the divergence between their results and ours. Analyses are by intention to treat.

The trial database, during the study phase, held and still holds the best (most accurate) address for each woman. These addresses did not include postal codes. When we noted in 1983 that there was an imbalance in all-cause mortality between the

Age at entry (years)	Intervention group		Control group	
	Number of women	Woman-years of follow-up	Number of women	Woman-years of follow-up
Cohort 1				
45–49	5777	78761	5594	75726
50–54	5878	78838	5168	68316
55–59	6109	79500	5749	73507
60–64	5162	64055	4831	58814
45–64	22926	301155	21342	276363
Cohort 2				
45–49	2495	29414	2381	28029
Cohort 3				
45–49	3207	31693	2292	22658

Table 1: **Trial population and woman-years of follow-up**

two trial groups, addresses of 20% of women from each general practice were given postal codes manually and 1981 census data derived from these postal codes were used to allocate an SES score to each practice.[4] From this score, practices were classified into three groups (high, medium, and low SES). This classification system (SES-1) has the strength that it was derived before any inspection of breast-cancer mortality but it did not succeed in eliminating the difference in other-cause mortality.[2,3] Postal codes have now been allocated by computer to the addresses of almost all women in the trial, allowing retrieval of the Carstairs index of deprivation[12] for the area of residence of each woman. The mean value of Carstairs index for general practices was used to derive a continuous variable (Carstairs score) for general practices.[5] The Carstairs score was used to assign practices into three groups (system SES-2) as before.

All analyses were repeated with adjustment for each of SES-1 (as used previously), and SES-2 and Carstairs score in the logistic regression model. In these adjusted analysis, we fitted terms for SES using the total trial population, even to report age-specific results. This approach assumes that the effect of SES is independent of age entry to the trial. That assumption has been verified; the interaction between SES and age at entry did not approach conventional levels of significance.

Results

Table 1 shows the number of women and woman-years at risk in cohort 1 by age at entry and for those aged 45–49 years at entry in cohorts 2 and 3. Women in cohorts 2 and 3 were generally younger at age of entry than those in the 45–49-year age-group in cohort 1 (mean age cohort 1, 47·4 years; cohort 2, 46·1 years; cohort 3, 45·8 years). When the cohorts are combined, they allow a direct comparison of 22 746 women who were younger than 50 years at entry (11 479 intervention and 10 267 control).

Women in the intervention group were offered an initial screening and up to six further screenings (three mammographic) for the initial cohort, four (with two mammographic) for cohort 2, and two (one mammographic) for cohort 3. 61·3% of women in the intervention group accepted the first invitation to screening, but attendance rates fell with time and were just over 50% during the final year of the study. Attendance at screening for women in the intervention group was strongly associated with SES (table 2). Younger women (45–49 years at entry) began trial screening under age 50 years but many continued to receive trial screening after 50 years. Of all trial screenings for these younger women 46·0% for the initial cohort, 79·0% for cohort 2, and 97·5% for cohort 3 were done when they were under 50 years. Little screening was available in Edinburgh during the study period (1978–88) and we believe that very few women in the control group arranged screening for themselves, although we have no way of confirming this assumption.

SES-2	Number of women from intervention group		Number of first-time attenders attending last screen
	Attending first screen	Attending last screen	
High	8212 (79%)	5747 (55%)	5747 (70%)
Medium	3961 (73%)	2632 (49%)	2632 (66%)
Low	2507 (70%)	1515 (42%)	1515 (60%)

Table 2: **Proportions of women attending for screening according to SES-2**

In cohort 1 followed up for 14 years there were 323 deaths in which breast cancer was the underlying cause. Analysis without adjustment for SES or with the previous method of adjustment (SES-1) showed differences of 13–16% between the intervention groups, which were not significant (table 3). Analysis of the unadjusted data by the TEDBC method gave almost identical point estimates.

However, adjustment by SES-2, in three categories or as a continuous variable (Carstairs score), gave point estimates for differences in breast-cancer mortality rates of 21–22%, which are of borderline significance (p=0·055 and 0·05, respectively). There was no significant interaction between age-group at entry and trial group (p=0·6), but we show results by age of entry within cohort 1 in table 3. All 95% CI are wide, but point estimates for all groups, except those of 50–54 years, are similar to the overall results; the difference in breast-cancer mortality between intervention and control groups for the youngest women (45–49 years) in the initial cohort was 30%.

When the late entrants (cohorts 2 and 3) were included in the 45–49-year age-group, the point estimate of the difference in breast-cancer mortality rates was 25%. Although the 95% CI was narrower, this difference in mortality is not significant. Cumulative breast-cancer mortality rates for the initial cohort at all ages and for women in all three cohorts entered at ages 45–49 years by trial group are shown in figure 2.

Age at entry (years)	Breast-cancer deaths				Rate ratio (95% CI)*		
	Intervention		Control		Unadjusted	Adjusted	
	n	Rate/10^5	n	Rate/10^5		SES-1	SES-2
Cohort 1							
45–49	27	3·43	33	4·36	0·78 (0·46–1·32)	0·73 (0·43–1·24)	0·70 (0·41–1·20)
50–54	44	5·64	35	5·16	1·09 (0·69–1·71)	1·03 (0·65–1·63)	0·99 (0·62–1·58)
55–59	43	5·49	55	7·56	0·71 (0·47–1·07)	0·68 (0·45–1·03)	0·65 (0·43–0·99)
60–64	42	6·67	44	7·55	0·87 (0·57–1·35)	0·83 (0·54–1·29)	0·80 (0·51–1·25)
45–64	156	5·18	167	6·04	0·87 (0·70–1·06)	0·82 (0·65–1·05)	0·79 (0·60–1·02)
All cohorts							
45–49	47	3·35	53	4·19	0·83 (0·54–1·27)	0·78 (0·50–1·21)	0·75 (0·48–1·18)

*Intervention group versus control group.

Table 3: **Breast-cancer mortality during 14 years of follow-up**

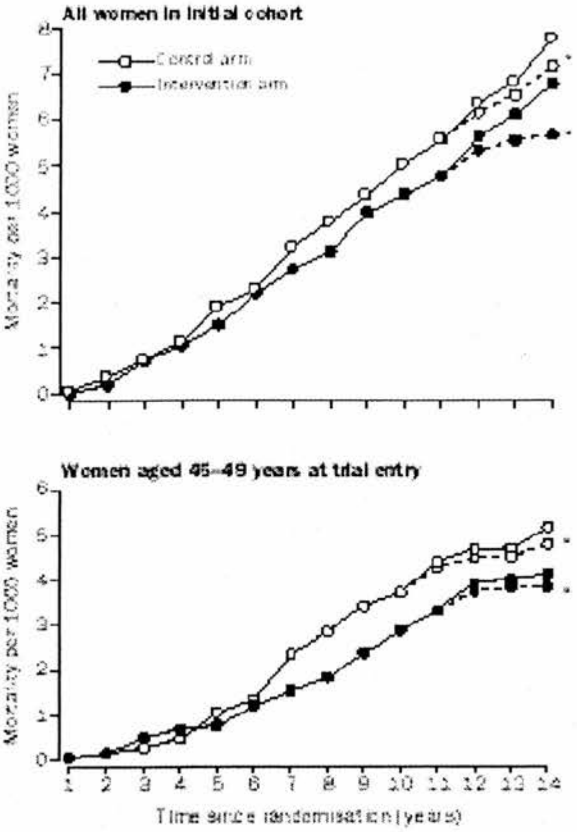


Figure 2: **Cumulative mortality from breast cancer (underlying cause of death) for all women in cohort 1 and for women aged 45–49 years at trial entry in all cohorts**
*Deaths following diagnoses up to study year 10.

When general practices were grouped by SES-2 there was a strong trend in the control group of mortality rates in cohort 1 and a weaker trend in the intervention group towards higher breast-cancer rates in those of higher SES (table 4). The benefit of screening seems to be concentrated in women of higher SES, but neither the interactions with SES-2 nor those with Carstairs score were significant.

Other-cause mortality rates were significantly lower in the intervention group than in the control group (rate ratio 0·84 [95% CI 0·79–0·90]) but after adjustment for Carstairs score (the continuous variable) this difference was no longer apparent (0·98 [0·92–1·04]).

In general, the results were not affected by censoring by study year of diagnosis. When censoring was applied to cases diagnosed after study year 10, the observed benefit in the intervention group was 29% (rate ratio 0·71 [0·53–0·95] if deaths after diagnoses after study year 10 are excluded). Figure 2 shows breast-cancer mortality with this censoring applied. Results were similar for the other censoring imposed for diagnosis around 3–4 years after the end of the study (study years 9–12, calendar years 1990–92).

SES-2	Intervention		Control		Rate ratio† (95% CI)
	n	Rate/10⁴	n	Rate/10⁴	
High	87	5·47	31	8·11	0·68 (0·44–1·03)
Medium	44	5·16	60	6·08	0·86 (0·58–1·27)
Low	25	4·39	76	5·45	0·78 (0·49–1·44)

*Breast cancer was the underlying cause of death. Rates are those per 10 000 woman-years at risk.
†Adjusted for age at survey entry, and, where appropriate, for cohort.

Table 4: **Breast-cancer mortality by SES-2 in cohort 1 (45–64 years)**

If deaths in the 45–49-year age-group were censored according to age at diagnosis, no benefit in the intervention group was observed when analyses were restricted to deaths of women whose breast cancers were diagnosed before they reached age 50 years (rate ratio 2·36 [0·79–7·08]); substantial benefit was seen when these deaths were excluded (0·47 [0·25–0·89]). The question of benefit from screening offered to women before they reach age 50 years (in addition to that from routine screening in their fifties) can be addressed, but with limited power, by comparison of breast-cancer mortality by trial group for younger women (entrants 45–46 years) in cohorts 2 and 3; for these women the rate ratios were 0·83 and 0·50, respectively (with wide 95% CI, 0·32–2·18 and 0·11–2·23).

Discussion

This report of the Edinburgh randomised trial of breast-cancer screening, with "before randomisation" consent and no offer of screening to the control group, includes nearly 70 000 woman-years of follow-up. Trials in Canada included a volunteer population all of whom, including the control group, received some form of intervention at entry to the trial[13] or throughout the study.[14] Because of recruitment of two later cohorts of women aged 45–49 years, the Edinburgh trial provides a substantial contribution to the total information available on the screening of younger women (more than 270 000 woman-years of follow-up). A limitation of this trial has been the imbalance between the intervention and control groups for other-cause mortality rates. New computer software has enabled us to access postal codes for all women, so we have been able to derive an improved estimate of general practice SES (SES-2).[5] The variation in SES explains the difference in other-cause mortality rates and thus allows best estimates of effects on breast-cancer mortality.

Application of SES-2 gave an overall point estimate for the difference between intervention and control groups in breast-cancer mortality rates in the Edinburgh trial at 14 years of follow-up as 21%; this estimate is of borderline significance (rate ratio 0·79 [95% CI 0·60–1·02]). If SES-2 is applied to our previously reported results at 10 years of follow-up, the difference in breast-cancer mortality rates is 24%, compared with the 18% first reported. This difference approaches significance (rate ratio 0·76 [0·55–1·06]).[5] These benefits do show some variation by age, with the largest benefits in women first screened in their late fifties and none in those first screened in their early fifties. Although this latter observation is consistent with results from some other trials,[15] we emphasise that the formal test of heterogeneity by age-group did not approach significance.

Breast cancers diagnosed after a suitable period from the end of the study clearly cannot have been influenced by trial screening, and the inclusion of deaths after these later diagnoses dilutes the comparison by trial group. In the HIP study, screening seemed to have no impact on diagnoses 3·0–3·5 years after the end of the study.[7] The first comparisons of the Swedish overview, based on short follow-up periods (mean 10 years) found only slight differences between follow-up (without censoring) and evaluation (with censoring) models, but later comparisons have shown increased estimates of benefit when censoring is imposed.[16] We found that estimated benefit was slightly larger when deaths from diagnoses more than 3–4 years after the end of the study were censored, with no evidence for attenuation of benefit at 14 years of follow-up. The HIP study[7] did show attenuation with, for example, deaths from diagnoses up to study year 7 showing benefits of 34·7%, 29·3%, and 22·2% at 7 years, 10 years, and 14 years of follow-up.

In view of the controversy over age at which screening should start, the results for women aged 45–49 years at entry are particularly important. Since there is no statistical evidence of heterogeneity between breast-cancer mortality benefit and age at entry, there is no reason to suppose that benefit is less for women first screened under 50 years than for older women. Furthermore, when age-specific results were derived, the estimated difference in breast-cancer mortality rates for those aged 45–49 years, although not significant by itself, is no less than in older entrants. In addition, there is no evidence that the benefit emerged later in these younger women. The 25% difference in mortality rate at 14 years of follow-up in women aged 45–49 years at entry agrees closely with most meta-analyses of randomised trials.[17]

There are the two critical components to the decision whether population screening for breast cancer should be available to women younger than 50 years. First, do women first screened when younger than 50 have lower breast-cancer mortality than those not so screened; and can the same benefit be achieved by screening from the age of 50 years? Even after a US National Institutes of Health consensus conference these issues remain controversial.[18-20] Although our numbers for such women are small, the Edinburgh trial findings make an important contribution to the first question.

The second question is more complex. Deaths classified by age at diagnosis require careful interpretation; for purposes of comparison with the HIP trial[7] we have presented results for younger entrants to the trial with diagnoses at 50 years or older and diagnoses at less than 50 years censored. Our results agree with those of the HIP trial that the benefit in women entering the trial before the age of 50 years is evident only in deaths occurring in their fifties and from cancers diagnosed after age 50 years. These observations could be artefacts with lead time advancing age at diagnosis to under 50 years in some women in the intervention group who, if unscreened, would have been diagnosed in their fifties. A proportion of the benefit (possibly 70%)[21] will be attributable to screening of women in their fifties. The design of the Edinburgh trial means that its analyses, based on randomisation, can address the second question for younger entrants to cohorts 2 and 3 (since almost all trial screening was done before age 50 years and women in both trial groups were eligible for Forrest screening), but the numbers available for analysis are small and the results equivocal. An observational study[22] on these cohorts and entrants at ages 45–49 years and 50–52 years to the initial cohort of the intervention group reported lower mortality rates for women in their fifties screened under 50 years.

The improved method of analysis and longer follow-up period have now shown a reduction of 21% in breast-cancer mortality for women aged 45–64 years at entry to the intervention group of the Edinburgh trial; this difference is of borderline significance. Consideration of the results at follow-up periods of 7 years and 10 years, and 14 years follow-up censored by study year of diagnosis indicates a benefit of 25–29% in a population offered regular screening (the steady state). The reduction is close to but smaller than the 30% expected when available data were restricted to HIP trial, the Swedish two-county trial, and case-control analyses of screening.[23] Compared with the Swedish two-county trial[24] pathological characteristics of cancers in screened women in Edinburgh showed that screening in Edinburgh (which used current mammographic technology) has not advanced the diagnosis sufficiently to influence histological grade despite reducing size and frequency of node involvement.[25] Although there is no reason to believe this finding is not typical of the UK, it may explain why mortality benefit has not been larger. Mammographic standards have certainly improved since the Edinburgh study but our data come from a research setting with a 2-year interval between screening. These data may not be applicable without reduction in benefit to screening done as part of routine health service with a 3-year interval. Only cautious optimism is appropriate. Subgroup analyses presented here and by the Swedish investigators,[15] and one Swedish trial of younger women,[26] are very promising for women younger than 50 years when first screened, but we believe that decisions on service screening in this age-group should await the results of specifically designed randomised trials (UK Age Trial and EUROTRIAL).

Contributors

Freda Alexander was responsible for leading the trial during its later stages, advised on the statistical analysis and study design, and drafted the paper; Tom Anderson was responsible for pathological aspects and assisted in the preparation of the paper; Helen Brown did most of the statistical analyses; Patrick Forrest provided advice on treatment and causes of death, and assisted in drafting the paper; Wilma Hepburn constructed the population register from which the trial population and the random allocation of women was derived, and was responsible for the flagging procedures; Alistair Kirkpatrick and Berenicé Muir were responsible for radiology; Robin Prescott advised on statistical analyses and assisted in drafting the paper; Alice Smith was data manager.

Acknowledgments

We thank other members of the project committee for advice and encouragement during the study: M M Andrew, J J K Best, C Brough, W Forbes, R Gruer, A Huggins, L J Kinlen, N B Loudon, W Lutz, U MacLean, M M Roberts, J Duncan, and I Sutherland. We wish to record the contribution by M Roberts, who was director of the Edinburgh breast screening clinic from 1979 until her death in 1989. We also thank M A Miller and B Moir for help in coordinating the funding; J Warner and other members of the staff of the Information and Statistics Division of the Scottish Health Service Common Services Agency, the staff at the General Register Office, Scotland, the Lothian Health Board Primary Health Care Division, and general practitioners in Edinburgh for their assistance; our colleagues in the screening clinic and the University Departments of Public Health Sciences, Pathology, and Surgery; and P Bisset for typing the paper.

The Edinburgh randomised trial was funded by the Cancer Research Campaign and the Chief Scientist's Office (CSO) of the Scottish Home and Health Department.

References

1 Roberts MM, Alexander FE, Anderson TJ, et al. The Edinburgh randomised trial of screening for breast cancer: description of method. *Br J Cancer* 1984; **50:** 1–6.

2 Roberts MM, Alexander FE, Anderson TJ, et al. Edinburgh trial of screening for breast cancer: mortality at seven years. *Lancet* 1990; **335:** 241–46.

3 Alexander FE, Anderson TJ, Brown HK, et al. The Edinburgh Randomised Trial of Breast Cancer Screening: results after 10 years of follow-up. *Br J Cancer* 1994; **70:** 542–48.

4 Alexander F, Roberts MM, Lutx W, Hepburn W. Randomisation by cluster and the problem of social class bias. *J Epidemiol Community Health* 1989; **93:** 29–36.

5 Alexander FE, Brown H, Prescott RJ. Improved classification of socio-economic status explains differences in all-cause mortality in a randomised trial of breast cancer screening. *J Epidemiol Biostat* 1998; **3:** 219–24.

6 Nyström L, Rutqvist LE, Wall S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993; **341:** 973–78.

7 Shapiro S, Venet W, Strax P, Venet L, Roser R. Ten year to fourteen year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982; **69:** 349–55.

8 UK Breast Cancer Detection Working Group. Trial of early detection of breast cancer: description of method. *Br J Cancer* 1981; **44:** 618–23.

9 UK Trial of Early Detection of Breast Cancer. 16-year mortality from breast cancer in the UK Trial of Early Detection of Breast Cancer. *Lancet* 1999; **353:** 1909–14.

10 OPCS mortality statistics: cause. Series DH2 No 11. London: HMSO, 1985.

11 Williams DA. Extra-binomial variation in logistic linear models. *ApplStat* 1982; **31:** 144–84.

12 Carstairs V, Morris R. Deprivation and health. *BMJ* 1989; **299:** 1462.

13 Miller AB, Baines CJ, To T, Wall C. Canadian national breast screening study: 1. Breast cancer detection and death rates among women aged 40–49 years. *Can Med Assoc J* 1992; **147:** 1459–76.

14 Miller AB, Baines CJ, To T, Wall C. Canadian national breast screening study: 2. Breast cancer detection and death rates among women aged 50–59 years. *Can Med Assoc J* 1992; **147:** 1477–88.

15 Andersson I, Aspegren K, Janzon K, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. *BMJ* 1988; **297:** 943–48.

16 Tabár L, Chen HH, Faberberg G, Duffy SW, Smith TC. Recent results from the Swedish Two-County trial: the effects of age, histological type, and mode of detection on the efficiency of breast screening. *J Natl Cancer Inst Monogr* 1997; **22:** 43–48.

17 Kerlikowske K, Grady D, Rubin SM, Sandrock C, Ernster VL. Efficacy of screening mammography. A meta-analysis. *JAMA* 1995; **273:** 149–54.

18 National Institutes of Health Consensus Development Panel. National Institutes of Health Consensus Development Conference Statement: breast cancer screening for women ages 40–49, January 21–23, 1997. *J Natl Cancer Inst* 1997; **89:** 1015–26.

19 Fletcher S. Breast cancer screening in women aged under 50. *BMJ* 1997; **314:** 764–65.

20 Eastman P. NCI adopts new mammography screening guidelines for women. *J Natl Cancer Inst* 1997; **89:** 538–40.

21 de Koning HJ, Boer R, Warmerdam PG, Beensterboer PMM, van der Maas PJ. Quantitative interpretation of age-specific mortality reduction from Swedish breast cancer-screening trials. *J Natl CancerInst* 1995; **87:** 1217–23.

22 Alexander FE. Edinburgh randomised trial of breast cancer screening. *J Natl Cancer Inst Monogr* 1997; **22:** 31–36.

23 Forrest APM. Breast cancer screening, report to the Health Ministers of England, Wales, Scotland and Northern Ireland by a working group. London: HMSO; 1987.

24 Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Gröntoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiol Clin North Am* 1992; **30:** 187–210.

25 Alexander FE, Anderson TJ, Hubbard AL. Screening status in relation to biological and chronological characteristics of breast cancers: a cross-sectional survey. *J Med Screen* 1997; **4:** 152–57.

26 Bjurstam N, Björeld J, Duffy SW, et al. The Gothenburg breast screening trial: first results on mortality, incidence, and mode of detection for women ages 39–49 years at randomisation. *Cancer* 1997; **80:** 2091–99.

Journal of Epidemiology and Biostatistics (1998) Vol. 3, No. 2, 219–224

Improved classification of socio-economic status explains differences in all-cause mortality in a randomised trial of breast cancer screening

FE ALEXANDER, HK BROWN and RJ PRESCOTT
Department of Public Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh, UK

Background The Edinburgh trial is one of only seven randomised controlled trials conducted to compare breast cancer mortality in women offered mammographic screening, with those having routine health care. Its importance has been diminished because cluster randomisation led to bias between the two arms: control women had higher all-cause mortality and lower socio-economic status (SES). Previous methods of quantifying SES could not explain the mortality bias.

Methods The Carstairs deprivation index has been used to derive an improved estimate of cluster SES. This has been validated by comparisons of non-breast cancer mortality using logistic regression and adjustment for SES. Similar analyses have been applied to breast cancer mortality.

Results Using the new estimate of cluster SES, the adjusted risk ratio for other-cause mortality for the intervention arm = 98 [95% confidence interval (CI) 0.92–1.14]. Adjusted risk ratios for breast cancer mortality at 7 and 10 years of follow-up are 0.79 and 0.77 respectively, with 95% CI = 0.54–1.18 and 0.56–1.06.

Conclusions The improved method of estimating cluster SES has overcome the problems which had previously been encountered. Analyses of breast cancer mortality using this adjustment should be free of bias. The estimated benefits in breast cancer mortality in women offered screening are larger than those previously reported at 7 and 10 years of follow-up.

Keywords breast cancer screening, mammography, socio-economic status, randomised controlled trial.

Introduction

The Edinburgh Randomised Trial of Breast Cancer Screening[1] (ERT) recruited women aged 45–64 years between 1978 and 1981. These women (n = 44 288) form the initial cohort. Younger women (45–49 years) also entered the trial in two further cohorts, 1982–3 (n = 4876), 1984–5 (n = 5499). A cluster randomisation process was based on an initial randomisation of 87 GP practices; women entering the trial were allocated by the status of their current GP into intervention and control arms.

Two reports of breast cancer mortality have been published: relating to 7 years[2] and 10 years[3] of follow-up. For both these follow-up periods we reported reductions of breast cancer mortality of 17–18%, which were not statistically significant.

Cluster randomisation is less effective than individual randomisation and, in this trial, it allowed bias between the two arms. Women in the control arm had higher all-cause mortality and lower socio-economic status[4] (SES). When the bias was first noted, addresses of a 20% sample of women were post-coded and several UK census variables related to SES were retrieved for the areas of residence in this sample. The first principle component of the census variables averaged over GP practices provided an estimate of practice SES[4]. Although this was significantly associated with all cause mortality, it failed to explain the differences between the two arms of the trial[2,3].

These biases have made interpretation of the results of the ERT difficult and diminished its perceived importance, although it is one of just seven randomised controlled trials of mammographic screening to have been conducted world-wide and one of only three outside Sweden. Adjustment for previous estimates of GP practice SES was applied in the 7-year analysis[2] and the estimated benefit in breast cancer mortality was somewhat increased as a result. This was not possible for the 10 year analysis because of a statistically significant interaction between the trial arm and SES[3].

Small area classification in the UK is now routinely based on well-validated indices of deprivation and/or socio-economic status derived from census data[5,6]; the index most widely used in Scotland is the Carstairs index[7]. Since computerised post-coding of addresses has become available, it is now possible to retrieve these indices for all women in the trial. The present report describes and justifies an improved method of quantify-

Correspondence to: FE Alexander, Department of Public Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh, EH8 9AG, UK.

ing SES using the Carstairs index and validates it in terms of other-cause mortality at 14 years of follow-up. Results at 7 and 10 years for breast cancer mortality, adjusted by this new measure, are also presented. Detailed examination of breast cancer mortality at 14 years follow-up will be reported separately elsewhere.

Methods

The trial population

Detailed methods have been described previously[1]. The geographical base was 87 general practices within the city of Edinburgh. All women aged 45–64 years registered with one of these and without a previous diagnosis of breast cancer were entered into the trial. The initial cohort was formed as the practices were entered sequentially between 1978 and 1981. During each of the 2 year periods 1982–3 (Cohort 2) and 1984–5 (Cohort 3) eligible women moving into one of these practices or attaining the age of 45 years and not already in the trial population were recruited. The majority of these were aged 45–49 years and only these have been included in the present report. The GP practice for each woman in the trial is the one with whom she was registered at randomisation.

Socio-economic status: initial classification

The trial database held, during the field work phase, and still holds, the 'best' address for each individual woman. These were not post-coded at the outset. When the potential bias was noted (1983), addresses of 20% of women from each GP practice were post-coded manually and 1981 census data were derived for the small census areas corresponding to these post-codes. The data included proportions of households with heads in social class I–II and IV–V, with no car and over-crowded and proportions of people seeking employment. The first principal component was taken as a socio-economic score (Princ_Comp) for each practice[4]. Low values of the score correspond to high SES. Princ_Comp showed a significant association with all-cause mortality and with mortality from several specific causes known to be associated with SES. From this score, practices were classified into three groups (high, medium, low), this is SES-1, corresponding to approximate tertiles of the distribution.

SES: logic of the new classification

- The Carstairs index of deprivation (DEPCAT[7]) for the small census area containing the post-code of the last recorded address is available for each woman and has been taken as the best measure of the SES of her area of residence.
- This is, however, *biased* since there was a systematic difference in time of recording of addresses between the two arms of the trial. Addresses of women screened, but not of other women, were routinely updated during the fieldwork period. In particular, a drift to lower SES may be anticipated for women who experience adverse health[8].
- We have an imperfect (since based on a 20% sample and using an *ad hoc* index), but unbiased, estimate of practice baseline SES, namely Princ_Comp.
- By averaging DEPCAT over GP practices we have another measure of practice SES, but subject to the bias noted above. However, after standardising the scales of this and Princ_Comp, the drift can be estimated and used to apply a bias correction to the measure based on DEPCAT.
- Since the drift is expected to be focused in women who have died (in the 14 years of follow-up now available) the procedure described above should exclude these women. We have also restricted our calculation to the initial cohort, as being more representative of the age range 45–64 years. The result (Carstairs_Score) is the best measure of practice SES now available to us.

SES: details of the new classification

Using current technology for computer address recognition, we have allocated post-codes to addresses of almost all (97%) women in the trial and using these, the Carstairs index of deprivation[7] for the area of residence of each woman has been retrieved. This index (DEPCAT) uses 1981 census data for the small areas in which the post-code is situated (low social class of head of household, households with no car, seeking employment, over-crowding). It has been extensively validated and is coded 1 (highest SES) to 7 (lowest SES).

The new measure (Carstairs_Score) has been derived in steps:

Step 1: the mean value (GP-DEPCAT) of the Carstairs index has been computer for each GP practice.

Step 2: the distributions of Princ_Comp and GP-DEPCAT across the women were compared and the former scaled so that the means and ranges of both were the same. This also stabilised the variance.

Step 3: the difference between (the scaled) Princ_Comp and GP-DEPCAT in the two arms of the trial were compared and the means for women in each trial arm calculated using only women in the initial cohort still alive at 14 years of follow-up. These mean differences were subtracted from GP-DEPCAT to yield Carstairs_Score. From this score, the practices were assigned to one of three groups (SES-2) with as nearly as possible the same number of women in each.

In order to investigate the SES drift for women who have died during follow-up, multiple regression analysis has been applied to the difference between individual DEPCAT and practice SES (Princ_Comp or Carstairs_ Score). The effect of trial arm has been examined after adjusting for age.

Mortality ascertainment and analysis

Women in the trials were flagged with the General Registry Office in Edinburgh in 1985 to obtain information on all-cause mortality. This ensures consistent follow-up for all women, even those who have left the city. The present report is restricted to women who were successfully flagged (98%). If breast cancer is mentioned on a death certificate then it is classified as underlying, or other cause, using WHO rules[9] as previously described[3]. All deaths of trial women for which breast cancer is not the underlying cause have been taken as 'other-cause' mortality.

For the statistical analysis of mortality, rates of death/10^4 women-years at risk have been computed and the ratio between the two arms of the trial calculated. As in previous analyses[2,3] a modified logistic regression procedure at the practice x age group level incorporated adjustment for extra-binomial variation[10] to respect the cluster randomization. These analyses were implemented in SAS using PROC LOGISTIC (with option Scale=Williams) and stratified by age at randomization 45–49, 50–54, 55–59, 60–64 years). When the later cohorts are included, further stratification by cohort has been added. Age-specific results have been derived from fitting interaction terms for age with trial arm, but including all women in the modelling process. Cumulative mortality curves are expressed as rates/10^4 women entered, but are adjusted to take account of women-years at risk.

In analyses with adjustment for SES in the logistic regression model, terms for SES (one of SES-1, SES2, Princ_Comp, Carstairs_Score) have been fitted to the total trial population (i.e. all age groups) even when age-specific results are reported. This means that the effect of SES has been taken to be independent of age at entry to the trial and this assumption has been verified by confirming that the interaction of SES and age-at-entry is not statistically significant.

Results

Table 1 shows, firstly, the number of women and women-years at risk using the longer follow-up now available, by 5 year age ranges at entry. The total numbers of other-cause deaths in the trial population are 6675 (initial cohort) and 309 (later cohorts). Analyses of other-cause mortality (Figure 1) show continued evidence of the difference between the two trial arms which has previously been reported.

Derivation of Carstairs Score

Both Princ_Comp and GP-DEPCAT were normally distributed over the trial population but Princ_Comp had a larger range (–2.35–2.40) and variance (1.14) than GP-DEPCAT (range: 1.46–5.42, variance: 0.76). In the sequel, Princ_Comp is replaced by a linear function with the same mean and range as GP-DEPCAT. Specifically, Princ_Comp was replaced by (Princ_Comp+2.35) × 0.80–1.96. The variance of Princ_Comp after transformation was 0.79.

Table 1 Trial population and women-years of follow-up

Cohort and age at entry (years)	Intervention arm		Control arm	
	Number of women	Women-years of follow-up	Number of women	Women-years of follow-up
Initial cohort				
45–49	5777	78 761	5594	75 726
50–54	5878	78 838	5168	68 316
55–59	6109	79 500	5749	73 507
60–64	5162	64 055	4831	58 814
45–64	22 926	301 155	21 342	276 363
1982–3 cohort				
45–49	2495	29 414	2381	28 029
1984–5 cohort				
45–49	3207	31 693	2292	22 658

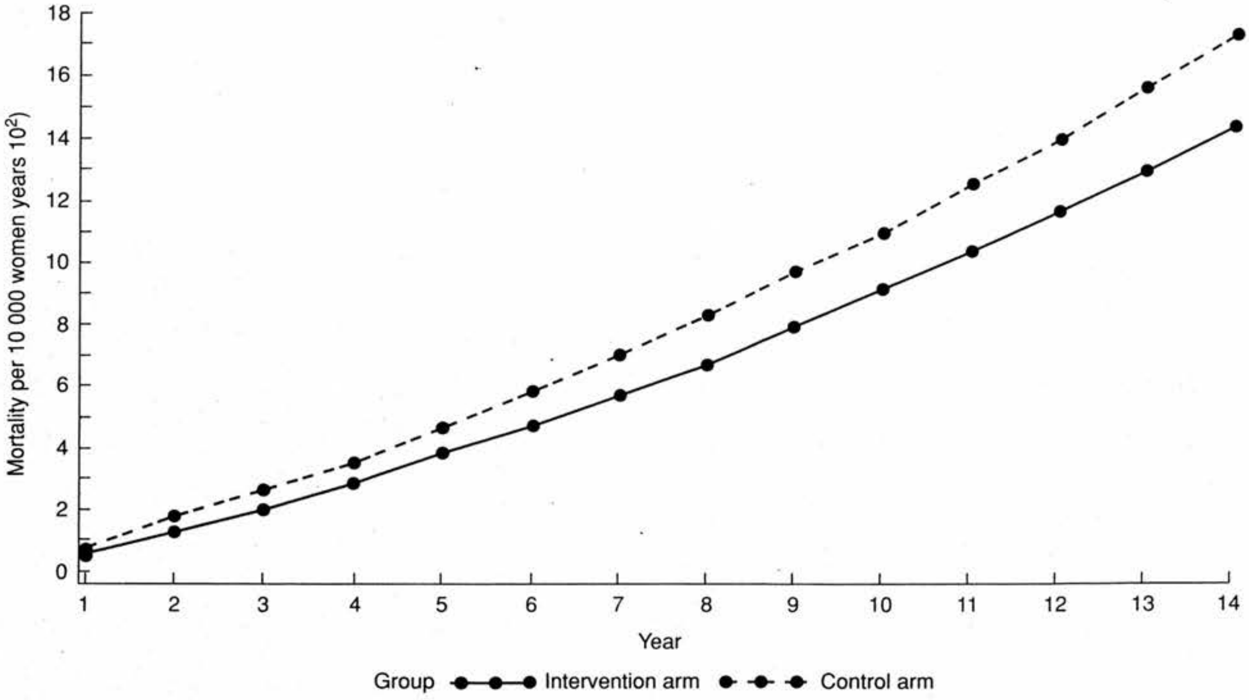


Fig. 1 Cumulative other-cause mortality (initial cohort only).

Mean differences between Princ_Comp and GP-DEPCAT for women in the two arms of the trial who were still alive at the end of the follow-up period are 0.05 (intervention arm) and –0.06 (control arm). These were taken to be the best available estimates of the bias in GP-DEPCAT caused by the timing of recording of addresses. Therefore, the new (bias corrected) measure of GP practice socio-economic status was:

$$\text{Carstairs\_Score} = \begin{cases} \text{GP\_DEPCAT-0.05 (intervention practices)} \\ \text{GP\_DEPCAT+0.06 (control practices)} \end{cases}$$

Both Pearson and Spearman rank correlation coefficients of Carstairs_Score with Princ_Comp were high when calculated for GP practices (0.94, 0.93) and individual women (0.92, 0.91).

In order to examine the evidence for a drift of SES for women who had died, the differences between the individual Carstairs index and Princ_Comp, Carstairs_Score were computed for women who had died (of causes other than breast cancer during the follow-up period). Use of either measure of practice SES (Table 2) shows a statistically significant difference by trial arm, with women in the intervention arm lower relative to their (initial) GP practice than those in the control arm.

Table 2 Comparisons by trial arm of individual Carstairs index (DEPCAT) with status of initial GP practice for women who have died[a] during follow-up

Measure	Regression coefficient for intervention arm[b]	95% CI
DEPCAT – Princ_Comp	0.125	0.05–0.20
DEPCAT – Carstairs_Score	0.231	0.15–0.31

[a] Of causes other than breast cancer

[b] Adjusted for age (four levels)

Figure 2 displays histograms of Carstairs_Score (using counts of all trial women) for intervention and control arms.

Validation: adjusted analyses of other-cause mortality (14 years follow-up)

Figure 1 displays cumulative other-cause mortality by trial arm; Table 3 shows comparisons by trial arm unadjusted for SES and with adjustment using each of SES-1, 2 and Princ_Comp Carstairs_Score. The unadjusted analyses and those adjusting for the original measures of practice SES reveal significantly lower other-cause mortality in the intervention arm. However,

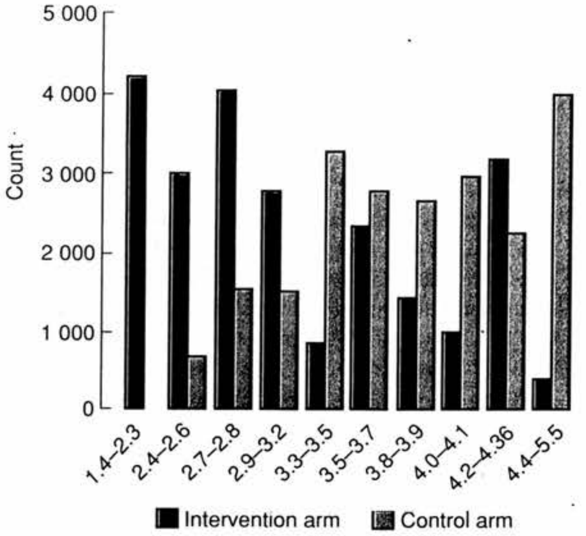


Fig. 2 Distribution of Carstairs_Score.

Table 3 Other-cause mortality: rate ratios for intervention arm compared with controls

	Age at entry and cohort	
Rate Ratios and 95% CI	**Initial cohort 45–64**	**All cohorts[a] 45–49**
Unadjusted[b]	0.84 (0.79–0.90)	0.82 (0.69–0.97)
Adjusted SES-1	0.89 (0.84–0.95)	0.88 (0.76–1.02)
Adjusted Princ_Comp	0.93 (0.88–0.98)	0.95 (0.87–1.04)
Adjusted SES-2	0.94 (0.88–1.00)	0.93 (0.80–1.08)
Adjusted Carstairs_Score	0.98 (0.92–1.04)	0.98 (0.85–1.13)

[a] 'Unadjusted analysis' here also includes terms for cohort
[b] 'Unadjusted analysis' includes terms for age at entry

adjustment for the new measure as a continuous variate effectively removes the difference by trial arm.

Adjusted analyses of breast cancer mortality (7, 10 years follow-up)

Table 4 gives breast cancer mortality results at 7 and 10 years of follow-up, with adjustment for Carstairs_Score. The estimated benefit in the intervention arm is larger than in previously published results. Risk ratios for the younger women have wide confidence intervals, but show no indication that benefit for them is either small or emerges later.

The Williams dispersion parameters were approximately equal to 1 (to within 0.1%) in all analyses, indicating that there was no over-dispersion in the data.

Table 4 Adjusted analyses of breast cancer mortality[a] (risk ratios[b] for intervention arm compared with controls)

Follow-up	**Group analysed**	
	Initial cohort (all ages)	**45–49 years[c]**
7 years	0.79 (0.52–1.21)	0.75 (0.32–1.77)
10 years	0.76 (0.55–1.06)	0.77 (0.44–1.33)

[a] Breast cancer as underlying cause
[b] With 95% confidence intervals
[c] Note that all women, including those entered 1982–1985, have the full follow-up period included in this analysis

Discussion

Randomisation in clinical, prophylactic and other trials is intended to ensure comparability of the groups compared for all relevant factors apart from the intervention being evaluated. The ability of randomisation to achieve equivalence of the groups is related to the number of randomisation units, rather than the number of individuals randomised. The ERT recruited and randomised a large number of women (over 50 000), but the number of randomisation units (clusters), was just 87. Optimal cluster randomisation requires the use of heterogeneous clusters which are similar to one another. GP practices are relatively homogeneous with respect to social and economic factors and dissimilar to one another. With hindsight, it is not too surprising that the randomisation process left differences between the trial arms in the ERT. Clearly, individual randomisation is preferable, but it was unacceptable to Edinburgh GPs in 1978. Despite strenuous efforts we have not, until now, succeeded in quantifying socio-economic status for the clusters in a way which can explain differences in all-cause (i.e. 'other'-cause) mortality.

Two alternative continuous variates have been considered here as methods of quantifying GP practice socio-economic status. Both are based on similar 1981 census data for the small areas into which individual addresses fall. The first (Princ_Comp) has the strength that it was derived before any data on breast cancer mortality were available; its weaknesses are, firstly, that it uses residences of just 20% of subjects and, secondly, that it has not proved able to explain differences of other-cause mortality between the two arms of the trial.

The second variate (Carstairs_Score) uses all addresses of living women and, in addition, uses the Carstairs index, which has been extensively validated and shown to be as powerful a predictor of mortality as personal SES[11] and to correlate highly with alternative UK indices[5,6]. Its weaknesses are, firstly, that it has been derived after examination of breast cancer mortality and, secondly, that there is a potential bias due to updating of addresses in the intervention arm. We have shown (Table 2) that this bias is a real problem for women who have died, since their SES has drifted downwards compared with their peers in the same GP practices at entry. These data provide warning to others, using socio-economic indicators based on addresses, that times of recording these should be carefully specified.

We have sought to minimise the problem of differential dates of recording addresses by restricting attention to living women and, in addition have made the best available bias correction. The best demonstration that this variate is optimal for classification of practice SES is its ability to explain almost totally the difference in other-cause mortality (Table 3). The categorical variables are worse than the continuous ones from which they derive. This is a natural consequence of the imbalance by arm at the extreme ends of the spectrum (Figure 2).

Use of new variate to adjust for SES in breast cancer mortality analyses (Table 4) gives estimates of benefit in the intervention arm which are larger than those reported previously and, at 10 years of follow-up, approach statistical significance. Future analyses of ERT data should use this method of adjustment confidently in the knowledge that, after it is made, no bias between the two trial arms remains in terms of non-breast cancer mortality.

References

1 Roberts MM, Alexander FE, Anderson TJ *et al*. The Edinburgh randomised trial of screening for breast cancer: description of method. *Br J Cancer* 1984;50:1–6.

2 Roberts MM, Alexander FE, Anderson TJ *et al*. Edinburgh trial of screening for breast cancer: mortality at seven years. *Lancet* 1990;335:241–6.

3 Alexander FE, Anderson TJ, Brown HK *et al*. The Edinburgh Randomised Trial of Breast Cancer Screening: results after 10 years of follow-up. *Br J Cancer* 1994;70:542–8.

4 Alexander F, Roberts MM, Lutz W, Hepburn W. Randomisation by cluster and the problem of social class bias. *J Epidemiol Commun Health* 1989;43:29–36.

5 Jarman B, Townsend P, Carstairs V. Deprivation indices. *Br J Med* 1991;303:523.

6 Morris R, Carstairs V. Which deprivation — a comparison of selected deprivation indices. *J Pub Health Med* 1997; 13:318–21.

7 Carstairs V, Morris R. Deprivation: explaining differences in mortality between Scotland, England and Wales. *Br Med J* 1989;199:886–9.

8 Townsend P, Davidson N. *Inequalities in health — the Black report*. Harmondsworth, UK: Penguin Books Ltd, 1982.

9 *OPCS Mortality Statistics: Cause*. Series DH2 No 11. London: HMSO, 1985.

10 Williams DA. Extra-binomial variation in logistic linear models. *Appl Stat* 1982;31:144–8.

11 Woodward M. Small area statistics as markers for personal social status in the Scottish Heart Health Survey. *J Epidemiol Commun Health* 1996;50:570–6.

Measurement of cyclic variation in ultrasonic integrated backscatter in conscious, unsedated, clinically normal dogs

Virginia Luis Fuentes, VetMB; Carmel M. Moran, BSc, PhD; Karsten Schober, DVM; Joanna Dukes McEwan, BVMS, MVM; Helen Brown, BA, MSc; George R. Sutherland, MBChB, MRCP; W. Norman McDicken, BSc, PhD

Objective—To assess the feasibility and repeatability of measuring ultrasonic integrated backscatter in unsedated conscious dogs, using a protocol previously validated in pigs with open thorax.

Animals—11 clinically normal conscious unsedated German Shorthair Pointers.

Procedure—A modified commercially available echocardiography system was used to record long-axis views of the heart. The radiofrequency data from 15 consecutive frames were digitized and analyzed. Regions of interest were chosen within the myocardium, and the ultrasonic integrated backscatter within each region was calculated in the time domain for each frame.

Results—Cyclic variation in integrated backscatter values was observed, with maximal values at end-diastole and minimal values at end-systole. Mean ± SD amplitude of cyclic variation was 5.81 ± 3.86 dB over all the regions chosen.

Conclusions—Results agreed with those obtained by other investigators working with dogs with open thorax and those with closed thorax while under general anesthesia. The analysis of the components of variance indicates that this is a consistent, reliable technique in conscious unsedated dogs.

Clinical Relevance—Integrated ultrasonic backscatter measurement provides a noninvasive means of tissue characterization. Use of this protocol reliably yields cyclic variation in integrated backscatter and could be applied clinically to dogs with myocardial disease. (*Am J Vet Res* 1997;58:1055–1059)

The interaction of ultrasound waves with biological tissue is strongly influenced by the physical characteristics of the tissue, and this interaction can be exploited to provide a noninvasive method of ultrasonic tissue characterization. Ultrasonic tissue characterization techniques have been documented to be capable of distinguishing normal from ischemic myocardium in studies of dogs with open thorax and in human patients,[1,2] and normal myocardium from that of human beings with dilated cardiomyopathy,[3] hypertrophic cardiomyopathy,[4] and diabetic myocardial disease.[5]

A number of methods have been described for the quantification of myocardial ultrasonic reflectivity by measuring small amplitude **radiofrequency (RF)** signals. However, techniques vary among laboratories, and there is no standard in vivo technique against which other techniques can be compared. **Integrated backscatter (IB)** is the most commonly measured in vivo variable, although a number of calculations have been used to derive IB, including frequency domain[6] and time domain techniques.[7] A cyclic variation with the phase of the cardiac cycle has been a consistent finding in myocardial IB, with maximal reflectivity generally corresponding to end-diastole. The magnitude of this cyclic variation has been found to vary with the myocardial site studied,[8] the angle of insonification, and left ventricular function. Most experimental studies have used dogs or pigs with open thorax. Of the few studies in which IB in dogs with closed thorax has been measured, anesthetized mixed breed dogs have generally been used.[9-11] To the authors' knowledge, reports of the measurement of IB in conscious, unsedated dogs do not exist.

The objective of the study reported here was to assess the feasibility of measuring ultrasonic IB from transthoracic views in conscious, unsedated dogs, using the protocol described by Moran et al[12] in pigs with open thorax. We also aimed to examine subject variation when a single breed of dog was used, and to measure intradog, interdog, and interobserver variability.

Materials and Methods

Dogs—Eleven German Shorthair Pointers (8 females, 3 males) were studied. Mean age was 6.0 (range, 1.5 to 12) years, and mean (± SD) body weight was 25.6 (± 3.5) kg. All dogs were considered to be healthy on the basis of historical findings and physical examination. All dogs were subjected to a routine two-dimensional, M-mode and Doppler echocardiographic examination to rule out any evidence of cardiac disease. All animals were used in accordance with the regulations and guidelines laid down in the UK Animals (Scientific Procedures) Act, 1986.

Imaging system and RF data acquisition—A modified commercially available digital echocardiography system[a] with

Received for publication Apr 26, 1996.
Manuscript passed review Oct 8, 1996.

From the Department of Veterinary Clinical Studies, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Summerhall (Luis Fuentes, Schober, McEwan); Medical Statistics Unit, Department of Public Health Sciences, University of Edinburgh Medical School, Teviot Place (Brown); the Department of Cardiology, Western General Hospital, Crewe Rd (Sutherland); and the Department of Medical Physics and Medical Engineering, University of Edinburgh, Royal Infirmary (Moran, McDicken), Edinburgh, United Kingdom. Dr. Luis Fuentes' present address is the Department of Veterinary Medicine and Surgery, College of Veterinary Medicine and Surgery, Clydesdale Hall, University of Missouri-Columbia, Columbia, MO 65211.

a 5 MHz phased array transducer (bandwidth, 2.5 to 7.5 MHz), and simultaneous ECG was used to acquire the two-dimensional echocardiographic images (Fig 1). Dogs were positioned in right lateral recumbency on a modified table, to allow placement of the transducer on the thoracic wall from below. Right parasternal long-axis views were used to obtain images of the left ventricle. Transmit power and time gain compensation controls were set at preadjusted levels for the studies, and the gain was adjusted at the start of each study, then was kept constant. The RF signals were digitized at 12 MHz to 16 bit. Fifteen consecutive frames of RF data were collected, and manually controlled so that the first frame occurred shortly before end-diastole. Each echo line of RF data consisted of a maximum of 4,096 samples, depending on the image depth. A single image consisted of 128 lines with up to 4,096 samples/line at 2 bytes/sample, occupying 1 megabyte (MB) of storage memory/frame. The RF acquisition system was equipped with a 16-MB memory board, of which 1 MB was reserved for control registers, so that a maximum of 15 complete frames of RF data could be acquired. The two-dimensional images were also stored on video, and hard copies were made of each frame.

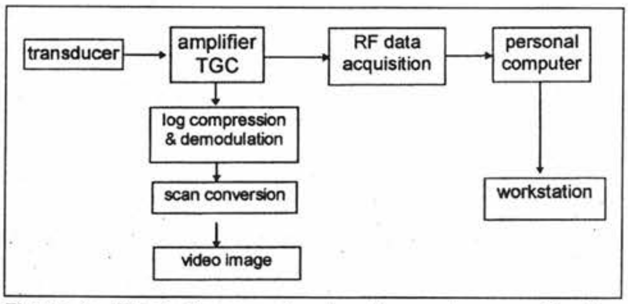


Figure 1—Block diagram showing the system used for the acquisition of radiofrequency (RF) data. TGC = time gain compensation.

The RF data were collected from each scan and downloaded onto a workstation[b] for signal processing. The RF data were rectified, low-pass filtered, logarithmically compressed, and scan converted before the reconstruction of two-dimensional images similar to the images obtained by the original echocardiography machine. Myocardial regions of interest in the reconstructed images were chosen within the left ventricular free wall and interventricular septum, using a mouse-driven cursor (Fig 2). Care was taken to avoid the specular echoes of the endocardial and epicardial surfaces. Regions of interest in the same myocardial area were tracked through each frame. These data were then related to the original uncompressed RF data, and the IB value was calculated. A set of calibration scans was also recorded, using a standard gray-scale test tissue phantom to compensate for the effects of transducer characteristics, time gain compensation, depth, and dynamic range settings of the ultrasound machine. The calibration scans were recorded for each depth and focus setting used during the in vivo scans, and the RF data were downloaded to the workstation in similar manner. The regions of interest chosen for each canine scan were then superimposed on the phantom scan RF data, and the IB was calculated according to the following formula:

$$\mathrm{IB} = \frac{\int_{t-\Delta t}^{t+\Delta t} |V(t)|^2 \, \delta t}{\int_{t-\Delta t}^{t+\Delta t} |P(t)|^2 \, \delta t}$$

where V(t) is the amplitude of the uncompressed RF signal from the myocardial region of interest, P(t) is the amplitude of the uncompressed RF signal from the phantom, δt is a small increment of time over the timegate, and 2Δt is the timegate over the region of interest.

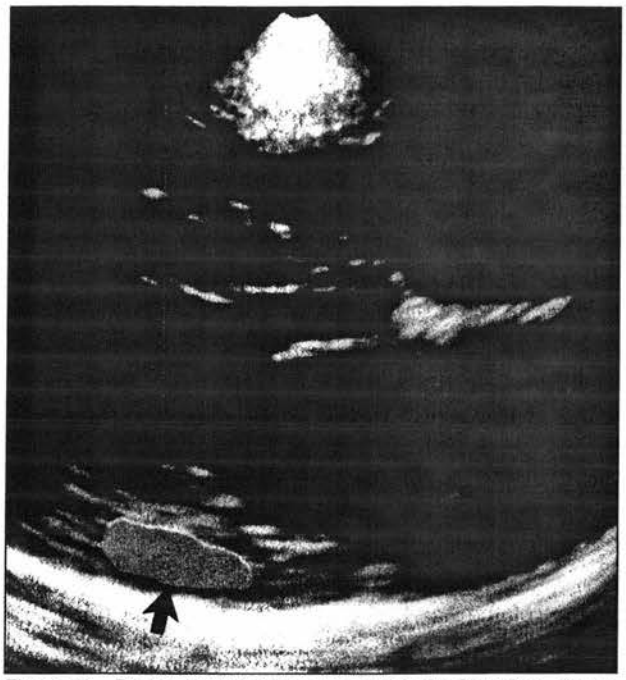


Figure 2—Reconstructed image on the workstation, showing region of interest chosen in the left ventricular free wall (arrow).

Protocol—All 11 dogs were scanned at least twice, and for each scan, values for IB were obtained for 3 myocardial regions over 15 frames (the basal left ventricular free wall, the apical left ventricular free wall, and the interventricular septum). In this manner, a set of IB values over 15 frames for each cardiac cycle was acquired in 86 sites. Ten scans from 8 dogs were analyzed by 3 independent observers over 3 regions, yielding 90 data sets comprising 15 frames for each data set. Three dogs were scanned 4 times, and 1 dog was scanned 3 times.

Analysis of data—The logarithm of each recorded IB value was multiplied by 10, and the mean value was subtracted to yield the variation about the mean, so that the mean value was designated as 0 dB. Using the concurrently recorded ECG, the 15 frames of RF data were synchronized for each scan by defining the end-diastolic frame as the frame recorded at the start of the R wave, and the end-systolic frame as the frame with the smallest left ventricular volume. The amplitude of cyclic variation was defined as the difference between the maximal and minimal values.

Statistical analysis—A random effects model was used to obtain variance component estimates for intradog, interdog, and interobserver variation within scans and within dogs. It was also used to obtain best least squares means and 95% confidence intervals. All calculations were carried out, using statistical analysis software.[c]

Results

Mean IB values for each frame from all the scans in all the regions of interest were obtained (Fig 3 top left), with the associated 95% confidence limits. Mean backscatter values and 95% confidence intervals for the individual regions (basal left ventricular free wall, apical left ventricular free wall, and interventricular septum)

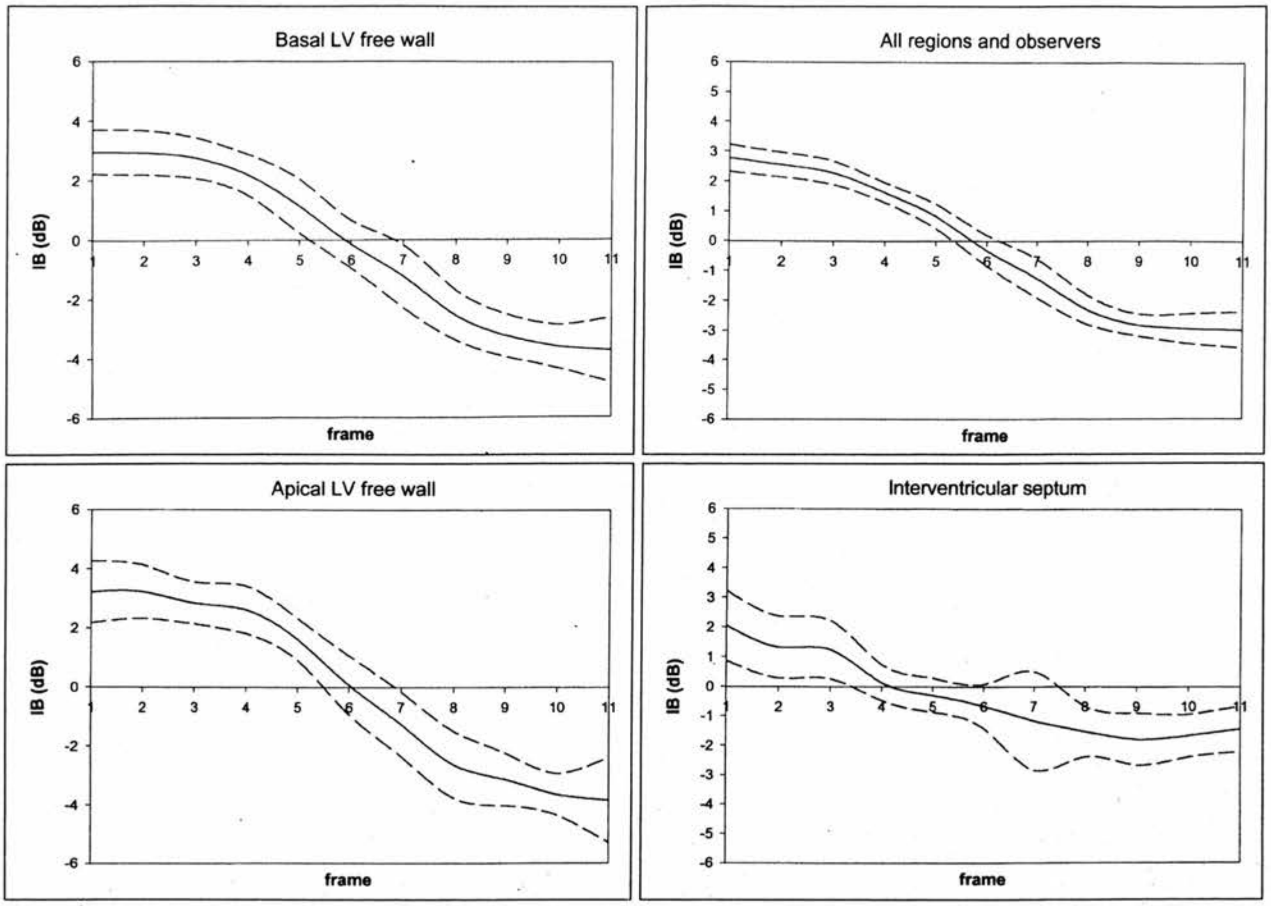


Figure 3—Mean integrated backscatter (IB) over 11 frames (from end-diastolic frame to end-systolic frame) for all scans and all observers. Mean IB levels for all 3 regions of interest and all 3 observers in all scans with 95% confidence intervals (top left). The same variables for the basal portion of the left ventricular (LV) free wall only (top right), apical portion of the LV free wall (bottom left), and the interventricular septum (bottom right).

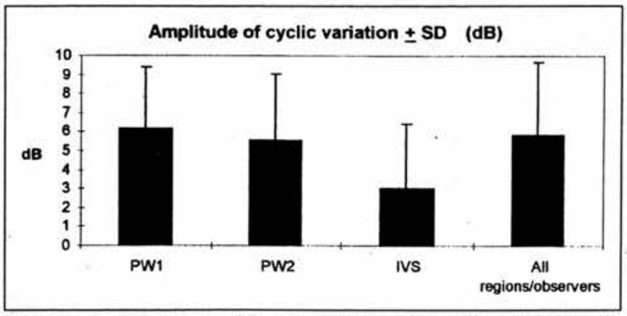


Figure 4—Mean ± SD amplitude of cyclic variation in basal region of the left ventricular free wall (PW1), apical region of left ventricular free wall (PW2), and interventricular septum (IVS).

also were determined (Fig 3 top right, bottom left, and bottom right). Amplitude of cyclic variation in each of the various regions was compared (Fig 4) and was greatest in the basal portion of the left ventricular free wall (PW1), with a mean ± SD value of 6.17 ± 3.19 dB. The apical portion of the left ventricular free wall was similar (5.55 ± 3.5 dB), but the interventricular septum had a lesser degree of cyclic variation (3.04 ± 3.4 dB). The variance components for interdog, intradog, and interobserver variation for each frame were compared (Table 1). The intradog variance ranged from 0.94 dB^2 for end-diastole to 1.08 dB^2 for end-systole, decreasing to approximately zero at mid-systole. Interdog variance was lowest at end diastole (0.45 dB^2) and end-systole (0.78 dB^2), and highest at mid-systole (1.75 dB^2). Variance components attributable to different observers (whether within each dog or within each scan) were smaller, reaching a maximum of 0.54 dB^2 for observer variance within each scan at end-systole. Other values are expressed as mean ± SD unless otherwise stated.

Discussion

Ultrasonic tissue characterization can provide unique insight into structural and functional changes in the myocardium, without recourse to invasive techniques, such as endomyocardial biopsy. Integrated backscatter measurement has been found to differentiate viable "stunned" ischemic myocardium from nonviable ischemic myocardium within 15 minutes of coronary occlusion in dogs before return to normal of any conventional echocardiographic variables, such as percentage of wall thickening.[1] Integrated backscatter has also been documented to differentiate hypertrophy attributable to athleticism from hypertrophic cardiomyopathy,[13] as well as hypertrophic cardiomyopathy from hypertrophy secondary to systemic hypertension.[14] A cyclic variation in IB, corresponding with the cardiac cycle, has been one of the most consistent findings in studies measuring IB in contracting myocardium.[9,15-23] The precise mechanism of this cyclic variation is unknown. It has been proposed that the intracellular and

Table 1—Variance components for integrated backscatter (IB) in all regions and all observers

	Integrated backscatter (dB)		Intradog variation		Observer variation within each dog		Observer variation within each scan		Interdog variation	
Frame	Mean	SEM	Variance (dB^2)	SD (dB)	Variance (dB^2)	SD (dB)	Variance (dB^2)	SD (dB)	Variance (dB^2)	SD (dB)
End-diastolic	2.78	0.23	0.94	0.97	0.00	0.00	0.43	0.66	0.43	0.66
2	2.56	0.21	0.85	0.92	0.12	0.35	0.22	0.47	0.65	0.81
3	2.28	0.2	0.49	0.70	0.21	0.46	0.07	0.26	0.96	0.98
4	1.68	0.17	0.32	0.57	0.04	0.20	0.27	0.52	0.44	0.66
5	0.83	0.21	0.00	0.00	0.10	0.32	0.31	0.56	0.95	0.97
6	−0.32	0.26	0.00	0.00	0.14	0.37	0.39	0.62	1.34	1.16
7	−1.25	0.33	0.00	0.00	0.20	0.45	0.48	0.69	1.75	1.32
8	−2.30	0.25	0.53	0.73	0.14	0.37	0.12	0.35	1.29	1.14
9	−2.83	0.19	0.30	0.55	0.32	0.57	0.09	0.30	0.81	0.90
10	−2.96	0.26	0.65	0.81	0.03	0.17	0.28	0.53	0.55	0.74
End-systolic	−2.99	0.31	1.08	1.04	0.13	0.36	0.54	0.73	0.78	0.88

extracellular elastic domains are cyclically altered in diastole and systole, thereby leading to dynamic changes in local acoustic impedance mismatch.[22] However, other research workers found that changes in IB occur in isotonic but not isometric contraction of canine papillary muscle, suggesting that tissue elastic properties may be unimportant.[24] In a study of backscatter variables in the frequency domain, using dogs with open thorax, manipulation of heart rate, preload, and mean arterial pressure did not affect the backscatter variables, and the authors suggested that the cyclic variation was caused by an effective change in shape, area, and orientation of the scatterers within the myocardium.[6] Results of a recent study have suggested that cyclic variation in IB may be directly related to sarcomere length.[25]

The IB values in this study varied in magnitude corresponding with the cardiac cycle, in agreement with the findings in most other studies. The maximal value was found at or near the end-diastolic frame, and the minimal value was seen at or near the end-systolic frame. The amplitude of cyclic variation was also similar to the findings of other groups, as well as to other studies using this protocol.[8,12] Other authors have also reported that the left ventricular free wall has greater amplitude of cyclic variation than that of the interventricular septum.[8]

Calibration of backscatter measurements by use of a tissue phantom has been described by Yao et al.[26] Using the ratio of the signal from the canine scans and the phantom scans eliminates the effects of the transducer characteristics. Attenuation of the RF signal by the thoracic wall in dogs with closed thorax will inevitably vary between systole and diastole, thus affecting the results. However, the IB values reported in this study correspond with those previously reported in dogs with open thorax, suggesting that these effects are probably minimal.[9-11]

A potential problem associated with the described protocol is the use of a 5-MHz probe when digitization of the RF signal is carried out at 12 MHz, thus introducing the possibility of signal aliasing. However, a fourth-order Butterworth filter was used within the echocardiography machine to limit aliasing and reduce noise. Calculations derived from this system suggest that $< 12\%$ of the signal power is likely to be derived from the aliased signal. Another limitation of this protocol is the inability to acquire more than 15 frames/scan. As with nearly any other cardiac variable, some beat-to-beat variation is inevitable, and this study assessed as few as 2 cardiac cycles in some dogs. Despite this, the SD value for intradog variation was approximately 1 dB for end-diastole and end-systole, and was negligible around early systole. The interdog variation was even less for end-diastole and end-systole (0.66 and 0.88 dB, respectively). Another disadvantage of this system is lack of an integrated timing reference, so that the simultaneous ECG was linked to the frames of the conventional scan, not to the specific RF data. The SD value for the end-diastolic to end-systolic amplitude of variation was large, although this may reflect the imprecise nature of the timing of individual frames, and does not take account of any phase delay of the cyclic variation.[9] However, the cyclic nature of the IB values was readily apparent (Fig 3), and this cyclic pattern would not be fully appreciated if analysis was limited to examination of end-diastolic and end-systolic frames only.

The advantage of the described protocol is the ability to collect RF data from the entire two-dimensional scan image. This provides freedom for the observer to choose specific regions of interest during subsequent off-line analysis, instead of limiting the data acquisition to regions predetermined at the time of scanning. The results suggest that the described protocol reliably indicates cyclic variation in integrated backscatter in normal, conscious dogs. Ultrasonic IB measurement offers the opportunity to measure myocardial tissue characteristics noninvasively, and has wide potential applications in the study of myocardial disease in small animals. Using the protocol described, this technique could be applied to clinical cases.

[a] ATL Ultramark 9 (UM9), Advanced Technology Laboratories, Bothell, Wash.

[b] Sparc System LX, Sun Microsystems Inc, Mountain View, Calif.

[c] PROC MIXED, SAS Institute Inc, Cary, NC.

References

1. Milunski MR, Mohr GA, Wear KA, et al. Early identification with ultrasoniç integrated backscatter of viable but stunned myocardium in dogs. *J Am Coll Cardiol* 1989;14:462–471.

2. Milunski MR, Mohr GA, Pérez JE, et al. Ultrasonic tissue characterization with integrated backscatter. Acute myocardial ischemia, reperfusion, and stunned myocardium in patients. *Circulation* 1989;80:491–503.

3. Vered Z, Barzilai B, Mohr GA, et al. Quantitative ultrasonic tissue characterisation with real-time integrated backscatter imaging in normal human subjects and in patients with dilated cardiomyopathy. *Circulation* 1987;76:1067–1073.

4. Lattanzi F, Spirito P, Picano E, et al. Quantitative assessment of ultrasonic myocardial reflectivity in hypertrophic cardiomyopathy. *J Am Coll Cardiol* 1991;17:1085–1090.

5. Pérez JE, McGill JB, Santiago JV, et al. Abnormal myocardial acoustic properties in diabetic patients and their correlation with the severity of disease. *J Am Coll Cardiol* 1992;19:1154–1162.

6. Sagar KB, Pelc LE, Rhyne TL, et al. Influence of heart rate, preload, afterload, and inotropic state on myocardial ultrasonic backscatter. *Circulation* 1988;77:478–483.

7. Thomas LJ III, Barzilai B, Pérez JE, et al. Quantitative realtime imaging of myocardium based on ultrasonic integrated backscatter. *IEEE Trans Ultrason Ferroelec Freq Control* 1989;36:466–470.

8. Lange A, Moran CM, Palka P, et al. The variation of integrated backscatter in human hearts in differing ultrasonic transthoracic views. *J Am Soc Echocard* 1995;8:830–838.

9. Mohr GA, Vered Z, Barzilai B, et al. Automated determination of the magnitude and time delay ("phase") of the cardiac cycle dependent variation of myocardial ultrasonic integrated backscatter. *Ultrason Imaging* 1989;11:245–259.

10. Barzilai B, Vered Z, Mohr GA, et al. Myocardial ultrasonic backscatter for characterization of ischemia and reperfusion: relationship to wall motion. *Ultrasound Med Biol* 1990;16:391–398.

11. Naito J, Masuyama T, Mano T, et al. Validation of transthoracic myocardial ultrasonic tissue characterization: comparison of transthoracic and open-chest measurements of integrated backscatter. *Ultrasound Med Biol* 1995;21:33–40.

12. Moran CM, Sutherland GR, Anderson T, et al. A comparison of methods used to calculate ultrasonic myocardial backscatter in the time domain. *Ultrasound Med Biol* 1994;20:543–550.

13. Lattanzi F, Di Bello V, Picano E, et al. Normal ultrasonic myocardial reflectivity in athletes with increased left ventricular mass. *Circulation* 1992;85:1828–1834.

14. Naito J, Masuyama T, Tanouchi J, et al. Analysis of transmural trend of myocardial integrated ultrasound backscatter for differentiation of hypertrophic cardiomyopathy and ventricular hypertrophy due to hypertension. *J Am Coll Cardiol* 1994;24:517–524.

15. Barzilai B, Madaras EI, Sobel BE, et al. Effects of myocardial contraction on ultrasonic backscatter before and after ischemia. *Am J Physiol* 1984;247:H478–H483.

16. Glueck RM, Mottley JG, Miller JG, et al. Effects of coronary artery occlusion and reperfusion on cardiac cycle-dependent variation of myocardial ultrasonic backscatter. *Circ Res* 1985;56:683–689.

17. Madaras EI, Barzilai B, Perez JE, et al. Changes in myocardial backscatter throughout the cardiac cycle. *Ultrason Imaging* 1983;5: 229–239.

18. Masuyama T, St Goar FG, Tye TL, et al. Ultrasonic tissue characterisation of human hypertrophied hearts in vivo with cardiac cycle-dependent variation in integrated backscatter. *Circulation* 1989; 80:925–934.

19. Milunski MR, Canter CE, Wickline SA, et al. Cardiac cycle-dependent variation of integrated backscatter is not distorted by abnormal myocardial wall motion in human subjects with paradoxical septal motion. *Ultrasound Med Biol* 1989;15:311–317.

20. Rhyne TL, Sagar KB, Wann SL, et al. The myocardial signature: absolute backscatter, cyclical variation, frequency variation, and statistics. *Ultrason Imaging* 1986;8:107–120.

21. Wear KA, Milunski MR, Wickline SA, et al. Contraction-related variation in frequency dependence of acoustic properties of canine myocardium. *J Acoust Soc Am* 1989;86:2067–2072.

22. Wickline SA, Lewis JT III, Miller JG, et al. A relationship between ultrasonic integrated backscatter and myocardial contractile function. *J Clin Invest* 1985;76:2151–2160.

23. Wickline SA, Thomas LJ III, Miller JG, et al. The dependence of myocardial ultrasonic integrated backscatter on contractile performance. *Circulation* 1985;72:183–192.

24. Wear KA, Shoup TA, Popp RL. Ultrasonic characterization of canine myocardium contraction. *IEEE Trans Ultrason Ferroelec Freq Control UFFC* 1986;33:347–353.

25. O'Brien PD, O'Brien WD, Rhyne TL, et al. Relation of ultrasonic backscatter and acoustic propagation properties to myofibrillar length and myocardial thickness. *Circulation* 1995;91:171–175.

26. Yao LX, Zagzebski JA, Madsen EL. Backscatter coefficient measurements using a reference phantom to extract depth-dependent instrumentation factors. *Ultrason Imaging* 1990;12:58–70.

Case-control study of sudden infant death syndrome in Scotland, 1992-5

Hazel Brooke, Angus Gibson, David Tappin, Helen Brown

Scottish Cot Death Trust, Royal Hospital for Sick Children, Glasgow G3 8SJ

Hazel Brooke, *executive director*

Angus Gibson, *chairman*

David Tappin, *senior lecturer in community paediatrics, University of Glasgow*

Medical Statistics Unit, University of Edinburgh Medical School, Edinburgh EH8 9AG

Helen Brown, *research associate*

Correspondence and requests for reprints to: Mrs Brooke.

BMJ 1997;314:1516–20

Abstract

Objective: To investigate the relation between routine infant care practices and the sudden infant death syndrome in Scotland.

Methods: National study of 201 infants dying of the sudden infant death syndrome (cases) and 276 controls by means of home interviews comparing methods of infant care and socioeconomic factors.

Results: Sleeping prone (odds ratio 6.96 (95% confidence interval 1.51 to 31.97)) and drug treatment in the previous week (odds ratio 2.33 (1.10 to 4.94)) were more common in the cases than controls on multivariate analysis. Smoking was confirmed as a significant risk factor (odds ratio for mother and father both smoking 5.19 (2.26 to 11.91)). The risk increased with the number of parents smoking ($P < 0.0001$), with the number of cigarettes smoked by mother or father ($P = 0.0001$), and with bed sharing ($P < 0.005$). A new finding was an increased risk of dying of the syndrome for infants who slept at night on a mattress previously used by another infant or adult (odds ratio 2.51 (1.39 to 4.52)). However, this increased risk was not established for mattresses totally covered by polyvinyl chloride.

Conclusions: Sleeping prone and parental smoking are confirmed as modifiable risk factors for the sudden infant death syndrome. Sleeping on an old mattress may be important but needs confirmation before recommendations can be made.

Introduction

There has been considerable interest in recent years about the role of infant care practices and environment in the sudden infant death syndrome. Previous studies showed the risk associated with sleeping prone,[1 2] and modification of this practice has been associated with a major reduction in the syndrome worldwide.[3] This improvement prompted our search for other modifiable risk factors. We report a four year case-control investigation of risk factors for the sudden infant death syndrome in Scotland from 1992 to 1995, when the rate of deaths from the syndrome fell from 1.1 to 0.7 per 1000 live births.

Methods

Population

The registrar general for Scotland reported to us all infant deaths occurring after the seventh day of life to the end of the first year. The computerised maternity record for each infant was provided by the information and statistics division of the Common Services Agency. In the case of deaths from the sudden infant death syndrome our office was also notified directly by the pathologist responsible for the necropsy. We defined the sudden infant death syndrome according to Beckwith as the sudden death of any infant or young child which is unexpected from the history and in which a thorough postmortem examination fails to show an adequate cause for death.[4]

Consistency in classification was sought by the use of a standard necropsy protocol with agreed diagnostic criteria.[5] In addition, all death certificates of infants aged 1 week to 1 year were scrutinised for possible misclassification of explained deaths. Overall, 201 out of 798 postperinatal infant deaths were diagnosed as the sudden infant death syndrome. Six other sudden deaths may have been misdiagnosed as bronchopneumonia and were not included in the study.

We identified two controls for each case of the syndrome—the births immediately before and after the index case in the same maternity unit. In this way controls were matched for age, season, and maternity unit. After permission had been obtained from the general practitioner a fieldworker employed by the study contacted the mother by letter, asking for her cooperation in completing a questionnaire during a home visit. All home visits were made within 21 days of the index case's death to minimise differences in age related circumstances between cases and controls. Questionnaires were completed on 147 cases out of a total of 201 reported and on 276 controls. The failure to acquire data on the remaining 54 cases was largely due to delay in notification by the pathologist or parents not being at home on at least two occasions, making a visit within 21 days of the death impossible. For 108 cases there were two controls, for 27 cases there was one control, and for 12 cases there were no controls. For 29 controls there was no case interview. The characteristics of the cases without an interview were compared with the characteristics of those with an interview and were similar in terms of maternal age, social class, and deprivation category.[6] There was a small difference in age at death of cases (the mean age of cases whose parent was interviewed was 15 weeks and that of those who were not was 18 weeks; $P = 0.04$).

Data collection

The questionnaire provided core medical and social data about the infant, as well as details of infant care practices in the home. Data were collected on routine childcare practices for cases and controls and on practice on the night of death for cases only.

The questionnaire was divided into six main categories: social and prenatal factors, feeding regimen, sleeping habits, sleeping environment, exposure to smoking, and illnesses.

Socioeconomic status was assessed by two methods. The first was assessment of deprivation on the basis of postal code in seven categories in ascending order of deprivation.[6] These categories take cognisance of overcrowding, male unemployment, low socioeconomic status, and a lack of a car. The second was the registrar general's social class from standard occupational classification.

The tog value (warmth rating) of both clothing and bedding was calculated with the scoring system supplied by the Shirley Institute, Manchester.[7] We used the average of day and night tog values in analyses. We took special note of the extra thermal implications of swaddling and how much of the body was swaddled.

Exposure to smoking was assessed in two ways. The first used the following ordinal scale: neither parent smoked, father only smoked, mother only smoked, mother and father both smoked. The second calculated a dose response by determining the total number of cigarettes (0, 1-9, 10-19, ≥20) smoked daily by mother, father, or other household member.

Exposure to old mattresses was assessed by asking parents if their infant routinely slept at night on a new mattress or on one previously used by another infant or an adult.

Data analysis

The primary analyses focused on routine childcare practices (see table 1), but routine practice and practice on the night of death were compared for cases (see table 2). The baseline comparison group always had the opposite definition—for example, maternal age <27 was compared with maternal age ≥27—unless otherwise stated in the footnotes to table 1.

Univariate analysis

Binary and categorical factors were analysed by conditional logistic regression analyses, which were fitted using the PROC PHREG program in SAS. This method allows for the matched nature of the data but is unable to use information on the unmatched cases and controls. Quantitative factors were analysed in random effects models fitted by the PROC MIXED program in SAS. Matched set was fitted as a random effect to allow for any correlations in the data due to matching. This method has the advantage that it uses information from unmatched cases and controls.

Over 100 factors were analysed, and about half of these were significant. However, many of the significant factors became non-significant when adjusted for obvious confounding factors or for socioeconomic factors. Results for the factors that remained significant after these adjustments are presented in table 1. Quantitative factors were categorised in this table so that all factors could be directly compared by odds ratios. Cut off points correspond to values expected to be most relevant on the basis of previous knowledge or as defined within other studies. All results in table 1 are based on comparisons between routine sleeping practices of cases and controls.

Table 1 Summary of risk factors for sudden infant death syndrome

	Proportion (%) of:					
Risk factor	**Cases**	**Controls**	**Univariate analysis odds ratio (95% CI)**	**P value**	**Multivariate analysis odds ratio (95% CI)***	**P value**
Exposure to smoking:						
Mother and father smoke	83/146 (57)	48/275 (17)	7.92 (4.32 to 14.55)		5.19 (2.26 to 11.91)	
Mother only smokes	32/146 (22)	45/275 (16)	5.26 (2.31 to 11.98)		5.05 (1.85 to 13.77)	
Father only smokes	11/146 (8)	45/275 (16)	1.72 (0.94 to 3.13)		2.12 (0.99 to 4.55)	
Neither parent smokes	20/146 (14)	137/275 (50)	1.00 (reference)	0.0001†	1.00 (reference)	0.0001†
Does not regularly change position during sleep	93/146 (64)	148/274 (54)	1.88 (1.01 to 2.50)	0.05	2.67 (1.42 to 4.99)	0.002
Old mattress used at night	91/145 (63)	108/274 (39)	2.50 (1.61 to 3.85)	0.0001	2.51 (1.39 to 4.52)	0.002
Maternal age <27	125/201 (62)	109/276 (39)	2.87 (1.85 to 4.45)	0.0001	2.37 (1.23 to 4.58)	0.01
Deprivation score of 7	46/195 (24)	25/266 (9)	9.59 (3.32 to 27.68)	0.0001	2.56 (1.20 to 5.49)	0.02
Drug treatment in previous week	49/131 (37)	43/186 (23)	1.99 (1.22 to 3.26)	0.02	2.33 (1.10 to 4.94)	0.03
Routine position put down to sleep[2]:						
Prone	13/146 (9)	5/275 (2)	5.37 (1.70 to 16.95)		6.96 (1.51 to 31.97)	
Side	75/146 (51)	104/275 (38)	1.58 (1.01 to 2.46)		1.51 (0.76 to 2.98)	
Variable	13/146 (9)	19/275 (7)	1.56 (0.71 to 3.42)		1.68 (0.52 to 5.42)	
Supine	45/146 (31)	147/275 (53)	1.00 (reference)	0.0001†	1.00 (reference)	0.04†
Has moved under bedclothes	35/146 (24)	35/274 (13)	2.48 (1.42 to 4.34)	0.001	2.18 (1.03 to 4.64)	0.04
Unmarried mother‡	130/200 (65)	87/276 (32)	4.22 (2.90 to 6.13)	0.0001	1.87 (1.00 to 3.48)	0.05
Social class IV or V§	88/201 (44)	53/276 (19)	2.55 (1.66 to 3.93)	0.0001	1.84 (0.99 to 3.43)	0.05
Male sex	138/201 (69)	138/275 (50)	1.84 (1.22 to 2.77)	0.004	1.76 (0.97 to 3.20)	0.06
Cot bumper not used routinely	104/146 (71)	154/274 (56)	2.00 (1.23 to 3.22)	0.005	1.74 (0.94 to 3.21)	0.08
Routinely sleeps with parents	11/146 (8)	6/275 (2)	3.92 (1.35 to 11.37)	0.01	2.90 (0.75 to 11.26)	>0.1
Any symptoms in previous week	113/147 (77)	162/275 (59)	2.16 (1.32 to 3.52)	0.002	1.58 (0.83 to 3.01)	>0.1
Gestation ≤36 weeks	44/196 (22)	17/276 (6)	4.49 (2.25 to 8.17)	0.0001	2.47 (0.67 to 9.12)	>0.1
Was usually swaddled in previous week	55/146 (38)	71/275 (26)	1.71 (1.06 to 2.76)	0.03	1.60 (0.77 to 3.33)	>0.1
Other infant death in family	21/146 (14)	14/276 (5)	2.76 (1.30 to 5.84)	0.008	2.45 (0.32 to 18.76)	>0.1
Usually sweaty on waking	68/146 (47)	77/275 (28)	1.93 (1.26 to 2.96)	0.003	1.27 (0.68 to 2.34)	>0.1
Tog value ≥10¶	45/147 (31)	50/275 (18)	1.98 (1.17 to 3.34)	0.01	1.10 (0.50 to 2.42)	>0.1
Mother left school aged ≤16	124/144 (86)	168/276 (61)	4.28 (2.41 to 7.62)	0.0001	1.07 (0.48 to 2.38)	>0.1
Not currently breast fed	137/147 (93)	209/275 (76)	4.35 (2.13 to 9.09)	0.0001	1.01 (0.39 to 2.71)	>0.1
Two or more previous live births	78/197 (40)	64/276 (23)	2.28 (1.46 to 3.56)	0.0003	0.99 (0.50 to 1.96)	>0.1
Birth weight <2.5 kg	42/197 (21)	19/276 (7)	3.79 (2.05 to 6.99)	0.0001	0.99 (0.29 to 3.38)	>0.1

*Odds ratios are adjusted for all other factors listed in table.

†For linear trend.

‡Includes mothers who were single, divorced, or cohabiting.

§Obtained from father's occupation or from mother's occupation if father's was unknown; includes 60 subjects for whom neither parent had an occupation to prevent their deletion in the multivariate model.

¶Average of day and night total tog values.

Multivariate analysis

A multivariate analysis was carried out to determine which factors were independently significant when adjusted for all other factors found to be important in the study. This was carried out using a conditional logistic regression model that included all of the factors listed in table 1.

Interactions

Interactions with factors of particular interest were tested in conditional logistic regression models.

Results

Frequencies, odds ratios, and P values for factors that remained significant after adjusting for obvious confounding factors and for socioeconomic factors are presented in table 1.

Nearly all of the people who smoked did so during and after pregnancy. Overall, 79% of mothers of cases (115/146) smoked compared with 34% of control mothers (93/275; univariate odds ratio 5.91 (95% confidence interval 3.61 to 9.68)). There was a dose relation, with the risk of the sudden infant death syndrome increasing with the number of cigarettes smoked by the mother ($P=0.0001$), father ($P=0.0001$), and other people in the household ($P=0.001$); each factor was analysed separately. The risk also increased significantly with exposure (table 1). The risk caused by maternal smoking increased when the infant shared a bed ($P<0.005$). Given that smoking is causal, the population attributable risk of smoking during and after pregnancy was 62%.

Sleeping prone remained a significant risk factor, although few infants in the control population were routinely placed prone (2%(5/275)); 9% of the index mothers (13/146) opted for this position routinely, resulting in an increased risk for their infants (table 1). At death 50 of the 147 infants were found prone (34%), though only 19 (13%) had been placed prone (table 2).

Table 2 Sleep practices routinely at night and at death in 147 babies who died of sudden infant death syndrome. Values are numbers of cases

	Routinely	At death
Placed prone	13	19
Found prone	16	50
Placed on side	75	78
Found on side	43	44
Placed variably	13	0
Found variably	21	0
Placed supine	45	47
Found supine	65	51
Total tog ≥10	45	40
Shared bed with parent	11	48
Duvet used	31	50
Pillow used	20	35
Swaddled	55	43

e 3 Old mattress, type of cover, and risk of sudden infant death syndrome uding routine bed sharers)

	Proportion (%) of:		Univariate odds ratio (95% CI)	Multivariate odds ratio (95% CI)*
of cover	Cases	Controls		
ttresses	78/130 (60)	101/264 (38)	2.27 (1.46 to 3.52)	2.46 (1.34 to 4.52)
nyl chloride throughout	24/36 (67)	30/50 (60)	0.92 (0.22 to 3.77)	0.44 (0.09 to 2.05)
	53/93 (57)	70/213 (33)	2.27 (1.29 to 4.00)	4.07 (1.90 to 8.68)

sted for all factors in table 1. Odds ratios remained virtually identical when infants bed sharing at were excluded.

Sleeping on the side was also a significant risk factor on univariate analysis (odds ratio 1.58 (1.01 to 2.46)). 51% of the index mothers (75/146) placed their babies on their sides, compared with 38% of controls (104/275). Sleeping on the side was the most labile sleeping position, but cases and controls routinely tended to move from sleeping on their side to sleeping supine if they moved at all rather than to sleeping prone. We noted that routinely only 44% (33/75) of index babies placed on their sides were found in a different position on wakening, compared with 68% (70/104) of controls (odds ratio 0.37 (0.20 to 0.68). Indeed, regardless of the position in which they were put down or found, controls were more likely than cases to change position regularly during sleep (odds ratio 0.53 (0.40 to 0.99)).

Mattresses used previously by at least one other infant or an adult seemed to place an infant at increased risk of the sudden infant death syndrome. The risk from routinely being on an old mattress at night also increased for infants who had undercovers with a lower tog value ($P=0.05$), who had colds ($P=0.02$), and who were off their feeds ($P=0.02$). No significant interactions occurred with smoking, sleeping place—for example, cot or pram—sleeping position, and birth order, so we could not draw any firm conclusions on the influence of these factors. There was no detectable increase in risk with old mattresses completely covered by polyvinyl chloride (table 3). The increase in risk was associated with both the so called combination mattresses, in which the bottom two thirds of the mattress is covered by polyvinyl chloride and the top third consists of ventilated foam covered by netting, and with cloth covered mattresses. We had insufficient data to investigate interaction of old mattresses with routine bed sharing. However, 34% of the index cases (48/142; table 2) were sleeping with parents at death—30% (42/142) in an adult bed and therefore on mattresses used by others.

We asked parents whether they had ever found that their baby had moved under the bedclothes. Overall, 24% of parents of cases (35/146) said that they had compared with 13% of controls (35/274), and this difference was significant on multivariate analysis (odds ratio 2.18 (1.03 to 4.64)) (table 1). On the day or night of death 13% (19/146) of the index cases were found under bedcovers.

Receiving any drug treatment in the week before death emerged as a strong risk factor in the multivariate analysis (table 1). After adjustment for symptoms in the previous week, consultation with a general practitioner, prematurity, and low birth weight, no individual drug was significant. We noted that the index cases were more likely to have had one or more of a range of symptoms and to have been seen by their general practitioner because of illness during the previous week. The symptoms with the highest odds ratios on univariate analysis were unusual sleepiness (odds ratio 2.61 (1.26 to 5.51)), snuffles (1.61 (1.07 to 2.44)), and sickness (1.69 (0.92 to 3.10)). The only symptom more common in controls was increased irritability.

Poverty was confirmed as a significant risk factor for the syndrome, the rate increasing with deprivation score, as shown in table 4. Low socioeconomic status (classes

Table 4 Association between deprivation score and sudden infant death syndrome

	No (%) of:	
Deprivation score	**Cases (n=195)**	**Controls (n=266)**
1	6 (3)	22 (8)
2	10 (5)	26 (10)
3	27 (14)	70 (26)
4	48 (25)	56 (21)
5	39 (20)	35 (13)
6	19 (10)	32 (12)
7	46 (24)	25 (9)

P<0.0001 for trend.

IV and V) was also significant even when adjusted for deprivation score in the multivariate model.

Factors significant on univariate but not on multivariate analysis were being male, sleeping on the side, non-routine use of cot bumper, routine sleeping with parent(s), any symptoms in previous week, gestation ≤36 weeks, usually being swaddled in previous week, previous infant death in family (sibling, half sibling, or first cousin), usually being sweaty on wakening, tog value of bedding and clothing ≥10, mother leaving school at ≤16, bottle feeding at time of death, two or more previous live births, birth weight <2500 g. However, their significance on univariate analysis makes them noteworthy. In particular, the finding of a high thermal score of bedding plus clothing is consistent with other published data.[8] A high thermal score seemed to be more risky for boys than for girls (P=0.03).

Examples of factors that were not significant on univariate analysis included the time between the last two pregnancies, twin birth, non-European mother, complementary feeding (combined breast and bottle feeding), age at introduction of solids, sleeping room, sleeping place (other than parents' bed), use of pillow, use of duvet, type of mattress covering, use of sheepskin, swaddling, heated sleeping room, and admission to hospital in week before death. Significantly more babies died on a Saturday or Sunday than would be expected by chance (42% (84/201), P<0.01); 11% (17/158) of deaths occurred when the infants were away from their usual place of residence.

The decrease in the rate of the sudden infant death syndrome over the four years was 0.4 per 1000 live births. We assessed whether attributable risks were likely to have changed during the study by testing whether factors had changed during the study in the control group. The following factors changed significantly: the number of parents who smoked was reduced (P=0.02), more infants were placed supine (P=0.01), and mothers were older (P=0.008). Thus, assuming constant relative risks, population attributable risks had decreased for parents' smoking, use of prone and side sleep positions, and younger mothers.

Discussion

Smoking

Parental smoking during and after pregnancy is a major, potentially modifiable, risk factor found in many other studies.[9 10] If only the father smoked the risk was almost significant. The finding of a dose response with the number of people smoking in the household adds weight to the possibility of smoking being causally related to the sudden infant death syndrome, as does the increased risk related to the number of cigarettes smoked. If smoking is causal two thirds of the cases of the syndrome might be avoided if mothers did not smoke during and after pregnancy. Health promotion initiatives to discourage young girls from starting to smoke and to help smokers reduce their habit are urgently required.[11 12]

Sleeping position and place

Although sleeping prone remains a strong risk factor, its low prevalence in the infant population in Scotland indicates that only a small percentage of deaths from the syndrome can now be attributed to this. Sleeping on the side was more risky than sleeping supine and since 38% of control infants were routinely placed this way, a significant number of deaths may be attributed to this sleep position. Sleeping on the back is the safest, so parents should be advised to use this position wherever possible.

The significance of failure to change position during sleep (table 1) supports observations by Schechtman *et al* that infants considered at increased risk of the syndrome show fewer spontaneous arousals from sleep and fewer movements during sleep than do control infants.[13]

The New Zealand cot death study found an increased risk for infants sleeping with a parent only if the mother smoked.[14] A subsequent report from California failed to confirm this risk,[15] but our data are consistent with the New Zealand findings. In addition, the greatly increased incidence of bed sharing in cases at death concerns us. We accept that routine bed sharing may be underreported, but it is difficult to believe that it could account for an increase from 8% routinely to 34% at death.

In Scotland most infants routinely share a room with their parent(s) at night—78% of index cases and 75% of controls. As nearly all the mothers in the study were European, the high prevalence of room sharing was not associated with an ethnic minority group, as noted in some studies.[16 17] In our study it was not a significant factor (odds ratio 1.20 (0.69 to 2.09) on univariate analysis. This differs from the data of Scragg *et al*, who found it to be protective.[18] We recognise that our data, in contrast to those of the New Zealand study,[14] were collected when the incidence of the syndrome and the rate of prone sleeping (2%) were low and the rate of room sharing was high (75%) in controls, making comparison difficult. On the basis of our results, however, we believe that advice on room sharing is not at present indicated in Scotland.

Mattresses and other factors

The results from the mattress analysis were unexpected. There is an increased risk of some kind, regardless of parity and social deprivation, for infants sleeping on mattresses previously used by others, although the risk was not established for mattresses completely covered by polyvinyl chloride. Our findings therefore lend no support to the hypothesis that household fungi interact with fire retardant chemicals in the plastic covering of cot mattresses and release toxic gases, which in turn cause sudden infant death.[19] The failure to establish risk with mattresses completely covered with polyvinyl chloride may be because they can be kept clean, regardless of age, while others cannot.

British Journal of Surgery 1996, 83, 997–1001

Factors affecting outcome of patients with impalpable breast cancer detected by breast screening

J. M. DIXON*†, O. RAVISEKAR†, M. CUNNINGHAM‡, E. D. C. ANDERSON†‡, T. J. ANDERSON§ and H. K. BROWN#

**University Department of Surgery, Royal Infirmary of Edinburgh, †Edinburgh Breast Unit, Western General Hospital, ‡South East of Scotland Breast Screening Centre, Springwell House, Ardmillan Terrace, and Departments of §Pathology and #Public Health Sciences, University of Edinburgh, Medical School, Teviot Place, Edinburgh, UK*

Correspondence to: Mr J. M. Dixon, University Department of Surgery, Royal Infirmary of Edinburgh, Edinburgh EH3 9YW, UK

Factors affecting completeness of excision and outcome, whether conservation or mastectomy, in 152 patients with localized impalpable breast cancer undergoing therapeutic needle-guided wide local excision were assessed by univariate and multivariate analyses using multiple logistic regression. Independent factors related to completeness of excision at the first operation were operator experience ($P = 0·0001$), and size of the lesion ($P = 0·005$). Factors related to outcome were operator experience ($P = 0·0003$), more experienced operators having a higher rate of breast conservation, and tumour size ($P = 0·0001$), larger lesions being more likely to be treated by mastectomy. Patients initially operated on by the two most experienced surgeons were more than four times less likely to undergo mastectomy than those whose initial wide local excision was performed by a less experienced surgeon.

Excision biopsy for impalpable breast cancer results in up to 60 per cent of patients having an incomplete excision[1,2]. It is now possible, with the use of stereotactic and ultrasonographically guided fine-needle aspiration cytology, to identify the majority of patients with impalpable malignant lesions before operation[3] and permit the first surgical procedure to be therapeutic. It is the policy in the authors' unit to perform a therapeutic needle-guided wide local excision for patients with localized impalpable lesions (4 cm or less) from which malignant cytology has been obtained[4]. If the initial excision is incomplete then patients are selected for either re-excision or mastectomy[5]. The aims of this study were to determine which factors affect the completeness of excision at initial operation and which influence whether the patient is ultimately treated by breast-conserving therapy or mastectomy.

Patients and methods

A total of 152 patients with localized (4 cm or less on mammography) screen-detected impalpable breast tumours diagnosed by stereotactic or ultrasonographically guided fine-needle aspiration cytology underwent a therapeutic wide local excision between 1 January 1991 and 30 June 1993. The notes of these patients were reviewed and data extracted on the factors listed in *Table 1*. These factors were then correlated with the outcome of surgery.

Patients were defined as having had a complete excision if the mammographic lesion had been completely excised and all resection margins were 1 mm or more clear of invasive and *in situ* disease on histological assessment.

Surgeons

Two surgeons performed almost 70 per cent of the operations and these have been designated as experienced operators (*Table 2*). One was a consultant and the other a senior registrar with a specific interest in breast disease who had been attached to the unit three times for a total of 3·5 years. In contrast, six of the seven inexperienced surgeons were attached to the unit for short periods of training. During their attachment to the breast unit, senior registrars and registrars were trained to perform wire localization biopsy in patients undergoing diagnostic excision. Thereafter, they were supervised during the first three wire-guided therapeutic wide local excisions for cancer and were then allowed to perform these procedures with no consultant present at the operating table, but available in the theatre suite.

Table 1 Factors studied

Nature of mammographic lesion
- Presence of microcalcification
- Type of microcalcification (branched or non-branched)
- Presence of mass lesion
- Distortion of architecture
- Combination of these features

Operator
- Experience (log of number of operations performed during the study)
- Experience (more than 20 operations) *versus* no experience (less than 15 operations) during the study

Method of localization
- Ultrasonography *versus* stereotactic

Pathological features
- Size of lesion (cm)
- Volume of excision
- Invasive cancer or *in situ* disease
- Type of invasive or *in situ* cancer
- Presence of extensive *in situ* component within invasive cancer
- Histological axillary node involvement

Operative procedure

All patients underwent wire-guided local excision. This was performed using stereotactic guidance in 121 patients and ultrasonographic guidance in 31. After removal of the specimen it was orientated with ligaclips and orientated specimen radiography was performed in an attempt to achieve complete excision at one operation as previously described[4]. The external surfaces of the specimen were painted with ink and the specimen placed on a trimming jig with the anterior surface uppermost. Serial slices (4 mm thick) were taken in a plane parallel to the chest wall and subjected to radiography. The appropriate areas

Paper accepted 7 September 1995

Table 2 Operators*

	No. of surgeons	Total no. of operations	Percentage of all operations	Grade of surgeon		
				Consultant	Senior registrar	Registrar
Experienced	2	106	71	1	1	—
Inexperienced	7	46	29	1	4	2

*Experienced surgeons performed more than 20 operations during the study period while inexperienced surgeons carried out less than 15

were blocked to include all six possible resection margins and all areas of radiological abnormality.

Protocol

A standard policy for re-excision and mastectomy was employed throughout the study period (*Table 3*).

Statistical analysis

Univariate analysis using logistic regression, χ^2 and Fisher's exact tests and multivariate analysis using multiple logistic regression were used to relate the factors in *Table 1* with completeness of excision at the first operation and outcome (breast conservation or mastectomy). Analyses were performed for the whole group (both invasive and *in situ* disease) and for the groups of patients with invasive and *in situ* disease separately.

Table 3 Unit protocol for re-excision and mastectomy

Re-excision
- One (focal) margin involvement and tumour either DCIS or invasive cancer and EIC-negative

Mastectomy
- Focal margin involvement and invasive cancer and EIC-positive
- More than one margin involved
- Margin involvement but impossible to determine extent or location of residual disease
- Multifocal invasive cancer
- DCIS more than 4 cm irrespective of margin involvement

DCIS, ductal carcinoma *in situ*; EIC, extensive *in situ* component

Results

Some 100 patients had invasive cancer and 52 *in situ* disease. The pathological size of the lesions ranged from 0·3 to 5 cm. Twelve patients had invasive tubular cancer and 22 of 100 with invasive cancer had an extensive *in situ* component. The volumes of excision ranged from 13 to 240 ml and seven of the 132 patients who had axillary nodes excised had involved axillary nodes.

A total of 122 lesions (80 per cent) were completely excised at the first operation, including five patients with ductal carcinoma *in situ* (DCIS) extending over a distance of more than 4 cm and two with multifocal disease. According to the protocol in *Table 3* these latter seven patients underwent mastectomy. Some 126 patients (83 per cent) were treated ultimately by conservation therapy and this included 115 who had an adequate excision at the first operation and 11 who underwent successful re-excision. Twenty-six patients (17 per cent) were treated by mastectomy.

Univariate analysis: all cancers

Of factors influencing the likelihood of complete excision and surgical outcome, operator experience and lesion size were significant in univariate analysis (*Tables 4* and *5*). The frequency of complete excision for each individual surgeon related to the number of operations carried out by that surgeon during the study is shown in *Fig. 1*. One of the inexperienced surgeons was a consultant who performed only a few procedures (*Table 2*). During their first operations the six inexperienced surgeons who were not consultants were supervised and 11 of these 18 excisions had clear histological margins compared with nine of 28 in later operations which were unsupervised. This difference was not statistically significant. A major

Table 4 Significant factors related to margin involvement in univariate analysis

	No. of patients	Excision (%)		*P*
		Complete	Incomplete	
Surgeon				
Experienced	106	102 (96)	4 (4)	0·0001
Inexperienced	46	20 (43)	26 (57)	
Microcalcification				
Yes	80	59 (74)	21 (26)	0·037
No	72	63 (88)	9 (12)	
Mass				
Yes	64	57 (89)	7 (11)	0·024
No	88	65 (74)	23 (26)	
EIC†				
Yes	22	15 (68)	7 (32)	0·022
No	78	69 (88)	9 (12)	
Mean(s.e.m.) size (cm)	—	1·40(0·09)	2·17(0·22)	0·0001

Values in parentheses are percentages. †Invasive cancers only. EIC, extensive *in situ* component

Table 5 Significant factors related to surgical outcome in univariate analysis*

	No. of operations	Breast conservation (%)			*P*
		One excision	Re-excision	Mast-ectomy	
Surgeon					
Experienced	106	96 (91)	1 (1)	9 (8)	0·0002
Inexperienced	46	19 (41)	10 (22)	17 (37)	
Mass					
Yes	88	59 (67)	9 (10)	20 (23)	0·031
No	64	56 (88)	2 (3)	6 (9)	
Invasive					
Yes	100	81 (81)	6 (6)	13 (13)	0·062
In situ					
Yes	52	34 (65)	5 (10)	13 (25)	

Values in parentheses are percentages. *Size also significant see *Table 6*

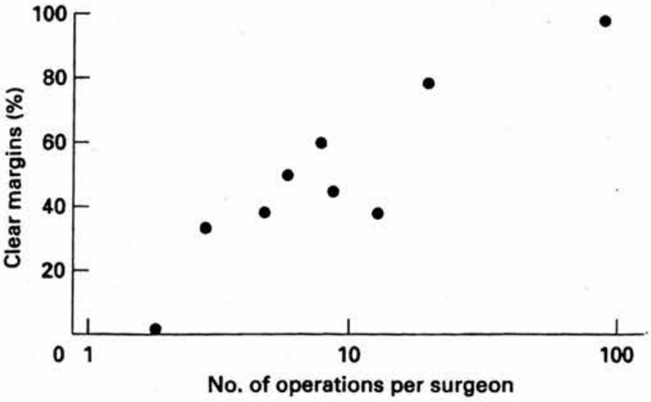


Fig. 1 Lesions completely excised compared with experience (plotted as the number of operations performed during the study – log scale)

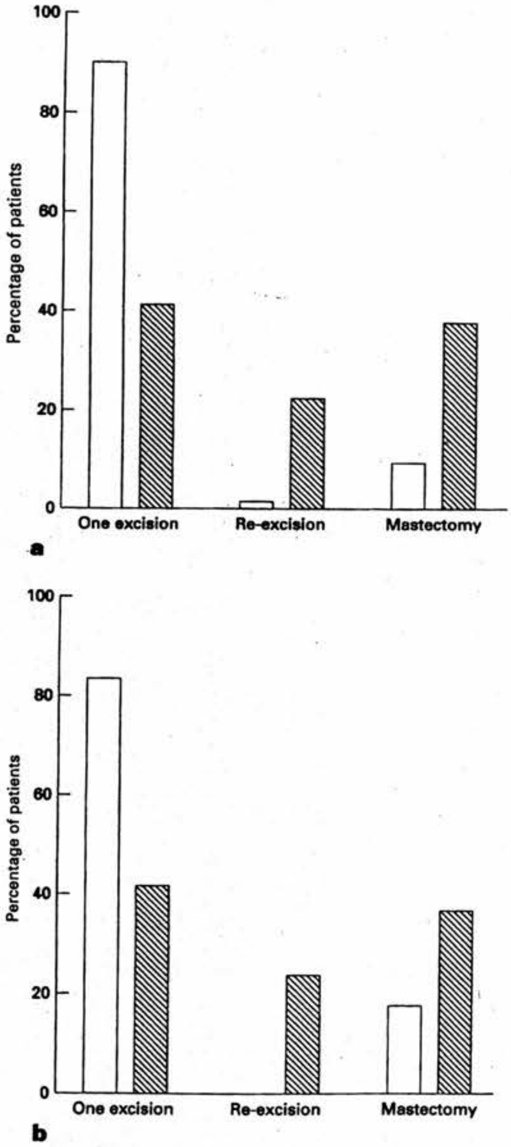


Fig. 2 Outcome in patients for **a** all cancers and **b** *in situ* cancers related to whether their initial operation was performed by an experienced (□) or inexperienced (▨) surgeon

problem during unsupervised operations was removal of areas of microcalcification in more than one piece, which made assessment of completeness of excision difficult. The three possible outcomes, namely complete excision at one operation, re-excision and mastectomy, according to whether the initial operation was performed by an 'experienced' or 'inexperienced' surgeon are given in *Fig. 2*. Patients who had their initial operation performed by an 'inexperienced' surgeon were greater than four times more likely to require a mastectomy than those whose initial operation was undertaken by a more experienced surgeon. The size of the lesion related to outcome is presented in *Table 6*: the mean size of the lesion increased as the extent of local surgery increased. The 12 patients with tubular cancer had a complete excision at the first operation but this was not significantly different from the rate of excision in those with invasive cancer of no special type.

Table 6 All cancers: size of lesion related to outcome

Outcome	No. of patients	Mean(s.e.m.) size (cm)
Single excision	115	1·25(0·06)
Re-excision	11	1·86(0·33)
Mastectomy	26	2·74(0·29)

Multivariate analysis: all cancer

Margins. The two factors retaining independent significance for completeness of excision were operator experience ($P = 0{\cdot}0001$), more experienced operators achieving a higher rate of complete excision, and size of lesion ($P = 0{\cdot}005$), larger tumours being less likely to be completely excised (*Table 7*).

Outcome. Factors related to outcome were operator experience ($P = 0{\cdot}0003$) and tumour size ($P = 0{\cdot}0001$). Patients were less likely to be treated by mastectomy if their initial operation was performed by a more experienced surgeon and larger tumours were more likely to be treated by mastectomy (*Table 7*). There were no significant relationships between other factors and outcome.

Invasive cancer

In multivariate analysis only operator experience ($P = 0{\cdot}005$) was related to completeness of excision (*Table 7*). Both operator experience ($P = 0{\cdot}006$) and tumour size ($P = 0{\cdot}028$) were identified as significant factors related to whether patients had breast conservation or mastectomy.

In situ *cancer*

Lesion extent and operator experience were the only independent significant factors predicting completeness of excision and outcome (*Table 7*). Experienced surgeons were more than twice as likely to excise an *in situ* tumour completely at one operation than less experienced surgeons (*Fig. 2*). The mean size of *in situ* lesions requiring mastectomy was over 3·5 cm (*Table 8*).

Table 7 Regression coefficients estimated by multivariate analysis

	Margins	Outcome
All cancer		
Operator experience	1·51(0·29)	0·98(0·24)
Tumour size	− 0·74(0·26)	− 1·42(0·32)
Invasive cancer		
Operator experience	1·25(0·33)	0·81(0·30)
Tumour size	− 0·57(0·42)	− 0·99(0·46)
In-situ cancer		
Operator experience	2·01(0·63)	1·17(0·49)
Tumour size	− 1·22(0·49)	− 1·87(0·61)

Values in parentheses are s.e.m.

Table 8 *In situ* cancer: size of lesion related to outcome

Outcome	No. of patients	Mean(s.e.m.) size (cm)
Single excision	34	1·53(0·15)
Re-excision	5	2·30(0·66)
Mastectomy	13	3·56(0·33)

Table 9 Reasons for mastectomy

Reason	No. of patients	Residual disease in mastectomy specimen
Invasive cancer		
Multifocal disease	3	0 of 3
Focal margin involvement and EIC-positive	3	1 of 3
Widespread margin involvement	2	1 of 2
Uncertainty about completeness of excision	5	3 of 5
In situ		
Extensive margin involvement	4	2 of 4
More than 4 cm margins clear	5	3 of 5
Uncertainty about completeness of excision	4	3 of 4
Total	26	13 of 26

EIC, extensive *in situ* component

Reasons for mastectomy

Seven patients underwent mastectomy after an apparently complete initial wide local excision because of DCIS extending over a distance of greater than 4 cm (five patients) or because of multifocal invasive cancer (two). The reasons for the remaining 19 patients having a mastectomy are outlined in *Table 9*. Of the 26 patients who underwent mastectomy, 13 were identified as having residual disease in the excised breast after mastectomy. Nine patients underwent mastectomy because of uncertainty over completeness of excision; all had their operations performed by unsupervised, 'inexperienced' surgeons. The finding of residual disease in six of nine of these patients justified mastectomy.

Discussion

The main role of screening is to reduce death from breast disease, but a secondary aim is to reduce morbidity. One way of achieving this is to treat more patients with breast cancer by breast-preserving surgery. Morbidity can also be reduced by identifying those patients with breast cancer before operation to permit the surgeon to perform a therapeutic excision as a first surgical procedure[4]. This study investigated those factors in patients with apparently localized impalpable breast cancer which affect whether the surgeon achieves a complete excision at initial operation and which influence whether patients are ultimately treated by breast conservation or mastectomy. The extent of the lesion was a major factor in determining the surgeon's ability to excise the lesion completely at the first operation. Disease extent also correlated with surgical outcome, with larger lesions being much more likely to require mastectomy.

The other major factor influencing completeness of excision and surgical outcome was experience of the surgeon who performed the initial wide local excision. Experienced surgeons were more than twice as likely to excise an impalpable tumour completely at one operation and patients whose initial operation was performed by an experienced surgeon were more than four times *less* likely to be eventually treated by mastectomy. The majority of inexperienced surgeons were those in training. They were trained according to an agreed protocol and were allowed to perform excision of impalpable tumours only when they were deemed to be competent. An adequate complete excision rate was achieved during the first few excisions when the trainees were supervised. When they started to perform operations alone, however, the rate of complete excision declined because they frequently removed the mammographic lesion in several pieces, which made assessment of completeness of excision and the site of any residual disease difficult. In such cases, the authors' protocol dictated that these patients should be treated with mastectomy; a policy that seems justified by the high rate of residual disease in the subsequent mastectomy specimens.

It was also policy to perform a mastectomy for areas of carcinoma *in situ* extending over 4 cm. As all mammographic lesions measured less than 4 cm, a pathological size of more than 4 cm indicated that the histological extent of disease was greater than the mammographic lesion and thus that in these patients was not a reliable marker of the extent of disease. The policy of advising mastectomy for these lesions is based on data from analysis of whole breast specimens[6]. The fact that residual disease was identified in three of five mastectomies performed on patients with DCIS over 4 cm which had apparently been completely excised justifies this policy and suggests that it would not be prudent to include DCIS lesions over 4 cm in size in the current UK randomized study of treatment of DCIS.

The presence of multifocal disease is usually considered an indication for mastectomy[7]. In the present series none of the patients with multifocal disease had residual disease in the removed breast. This suggests that breast conservation therapy may be appropriate in patients with localized multifocal disease[8]. The presence of an extensive *in situ* component (EIC) within an invasive cancer is a risk factor for residual disease in the breast after wide excision[9]. In the present series patients with invasive cancer which was EIC positive were significantly less likely to have an initial complete excision than those with tumours which were EIC negative ($P = 0·022$, *Table 4*). However, only one of three patients who had focal margin involvement and an EIC-positive tumour had residual disease in the breast at mastectomy but the numbers are too small to draw any firm conclusions.

The findings of the present study have important implications for the training of surgeons in the removal of screen-detected impalpable tumours. The aim during the study period was to train all senior registrars, and some registrars, attached to the breast unit to remove impalpable breast tumours safely. They started by performing wire localization biopsy under supervision in patients undergoing diagnostic excision. After performing three or four biopsy procedures they proceeded to supervise wire-guided therapeutic excision for breast cancer. When considered to be competent, they were permitted to perform these procedures without an experienced surgeon supervising the operation. This study demonstrates that there is a long learning curve to this procedure and that it is not until a trainee has performed 10–15 of these operations that he/she becomes competent.

References

1 Campbell ID, Royle GT, Coddington R *et al.* Technique and results of localization biopsy in a breast screening programme. *Br J Surg* 1991; **78**: 1113–15.

2 Aitken RJ, MacDonald HL, Kirkpatrick AE, Anderson TJ, Chetty U, Forrest APM. Outcome of surgery for non-palpable mammographic abnormalities. *Br J Surg* 1990; **77**: 673–6.

3 Azavedo E, Svane G, Auer G. Stereotactic fine-needle biopsy in 2594 mammographically detected non-palpable lesions. *Lancet* 1989; i: 1033–6.

4 Dixon JM, Ravi Sekar O, Walsh J, Paterson D, Anderson TJ. Specimen-orientated radiography helps define excision margins of malignant lesions detected by breast screening. *Br J Surg* 1993; **80**: 1001–2.

5 Schnitt SJ, Connolly JL, Harris JR, Hellman S, Cohen RB. Pathologic predictors of early local recurrence in stage I and II breast cancer treated by primary radiation therapy. *Cancer* 1984; **53**: 1049–57.

6 Holland R, Hendricks JHCL, Verbeek ALM, Mravunac M, Schuurmans Stekhoven JH. Extent, distribution, and mammographic/histological correlations of breast ductal carcinoma *in situ*. *Lancet* 1990; **335**: 519–22.

7 NIH Consensus Conference. Treatment of early-stage breast cancer. *J A M A* 1991; **265**: 391–5.

8 Kurtz JM, Jacquemier J, Amalric R *et al.* Breast-conserving therapy for macroscopically multiple cancers. *Ann Surg* 1990; **212**: 38–44.

9 Schnitt SJ, Connolly JL, Khettry U *et al.* Pathologic findings on re-excision of the primary site in breast cancer patients considered for treatment by primary radiation therapy. *Cancer* 1987; **59**: 675–81.

heart rates also declined in a similar manner to the first study. The relationship between the reduction in exercise heart rates and the increase in threshold for breathlessness was statistically significant ($P < 0{\cdot}001$).

Over the same time period, a further thirteen subjects (age 31 ± 10 years, mean ± S.D.) rebreathed CO$_2$ in oxygen utilizing the Read rebreathing procedure. Neither the slope nor the intercept of the VAS/VE relationship showed any consistent effect of repetition. Heart rate data were available on five subjects. There was no significant change (Student's t test) either during the actual test or between test days, whether one looked at the heart rates at a fixed index (7 % FeCO$_2$), the slope of the fC/FeCO$_2$ relationship, or at the beginning or end of the experiment.

The underlying mechanism for the observed changes in the threshold response could include an alteration in the criteria used by the subjects to estimate the magnitude of the sensation, and physiological effects resulting from repetitive exposure to the exercise i.e. a training response. The data from these experiments suggest that at least some of the change is associated with a cardiovascular training effect; the rebreathing study does not support the concept of a change in the assessment criteria as a result of repetitive exposure.

REFERENCE

Reed, J.W. & Subhan, M.M.F. (1994). *J. Physiol.* **480.P**, 54*P*.

Maturational changes in the respiratory rhythm of the mouse *in vitro* and *in vivo*

Julian F.R. Paton and Diethelm W. Richter

II Physiologisches Institut, University of Gottingen, Humboldtallee 23, D-37073, Germany

The effects of raising body or hypothalamic temperature on ventilation in the anaesthetized rat

M.J. Parkes*† and M.C. Harris†

**School of Sport & Exercise Sciences and † Department of Physiology, University of Birmingham, Birmingham B15 2TT*

There is evidence for a relationship between raised body temperature and apnoea in the neonatal period (Stanton, 1984). Since the adult rat makes certain respiratory responses which resemble those of the human neonate (Neylon & Marshall, 1991), it may be useful to study this relationship in the rat. Little, however, is known about the effects of raising body temperature on ventilation in this species (Gordon, 1990).

Adult rats therefore were anaesthetized with Hypnorm (1 ml kg^{-1} I.P.) and diazepam (1 ml kg^{-1} I.P.). The depth of anaesthesia was monitored carefully and α-chloralose (5 mg doses I.V.) was given if pinching the tail produced a change in blood pressure or heart rate. Oxygen-enriched air was breathed spontaneously through a tracheal cannula. End-tidal P_{CO_2}, heart rate, blood pressure, rectal and hypothalamic temperature were measured continuously.

In seven rats heat was applied over a 20–46 min period via a heating blanket and an infrared lamp. The frequency of spontaneous breathing was 72 ± 3 breaths min^{-1} (mean ± S.E.M.) at a hypothalamic temperature of 36·5 °C and increased to 96 ± 5 breaths min^{-1} at 39 °C ($P < 0{\cdot}001$, one-tail paired t test, $n = 6$). Heart rate increased from 447 ± 31 beats min^{-1} at 36·5 °C to 490 ± 21 beats min^{-1} at 39 °C ($P < 0{\cdot}002$, $n = 5$). There was no significant change in end-tidal or arterial P_{CO_2} (46 ± 1 mmHg at 36·5 °C and 46 ± 3 mmHg at 39 °C, $P > 0{\cdot}05$, $n = 5$). P_{CO_2} fell, however, by 1–5 mmHg in three rats, and rose by 4–9 mmHg in two rats. There was no obvious difference between rats which could explain why P_{CO_2} rose in some and fell in others.

In two rats localized heating in the pre-optic area (POA), on the right side, was attempted by passing a 10 MHz current for 60 s through a bipolar electrode. Tissue temperature 1 mm lateral to the electrode was measured and the position of the warming electrode confirmed histologically at post-mortem. In one rat breathing frequency was 61 breaths min^{-1} in the 85 s period preceding the onset of heating and increased to 93 breaths min^{-1} during the 85 s period following the onset of heating. In the other rat heating outside the POA produced no increase in breathing frequency.

These results show that in the anaesthetized rat raising body temperature or localized heating in the POA initially increases the frequency of spontaneous breathing. The rat therefore responds in a similar manner to other mammals and is suitable for more detailed investigation of the relationship between raised temperature and episodes of apnoea.

REFERENCES

Gordon, C.J. (1990). *Physiology & Behavior* **47**, 963–991.
Neylon, M. & Marshall, J.M. (1991). *J. Physiol.* **440**, 529–545.
Stanton, A.N. (1984). *Lancet* **ii**, 1189–1201.

Stability of breathing patterns in men, rats and rabbits

M. Dallak, Xiujie Luan*, H. Brown and L. Pirie

*Department of Physiology, University of Edinburgh, Edinburgh and *Bethune University of Medical Sciences, Changchun, China*

We intend to compare breathing pattern in conscious emphysematous and normal rabbits. To estimate the minimum number of emphysematous rabbits required to give a statistically valid result we need to know the variability of the pattern of breathing in normal rabbits.

We can find no record of such a measurement in the literature, although rabbits are said to have a highly labile pattern of breathing. The reproducibility of pulmonary mechanics has been measured in cattle (Galivan & McDonell, 1988) and breathing pattern of human beings has been measured in single instances (Tobin *et al.* 1983) and over time (Benchetrit *et al.* 1989).

We have adapted the whole body plethysmography method of Bartlett & Tenney (1970) developed from the method of Drorbaugh & Fenn (1955) to measure the breathing pattern of four rabbits and four rats on four separate days. The plethysmograph consists of a chamber through which a stream of air enters from a pump and leaves via a narrow tube. This tube offers high impedance to flow at the frequency of breathing of rats and rabbits. The breathing of the occupant of the chamber is therefore accurately reflected by small changes in pressure in the chamber. To precisely measure tidal volume the inlet and outlet of the chamber was closed for a few seconds. The breathing pattern of four human subjects was measured on four separate days, using an ultrasonic pneumotachograph (FIP Instruments, Field Road, Huntingdon, Cambridge). One hundred consecutive breaths were measured in terms of their inspiratory duration (t_I) expiratory duration (t_E) and tidal volume (V_T).

Variability of pattern was calculated as components of variance using the commercial analysis program SAS (SAS Institute Inc., SAS Circle, Box 8000 Cary, NC 27512–8000).

We calculate that to detect with 80% certainty a 10% change in the mean values of inspiration, expiration and tidal volume with the experimental protocol and species we have used, would require twelve men, nineteen rabbits, or eight rats respectively.

Supported by the Norman Salvesen Emphysema Research Trust.

REFERENCES

Bartlett, D. Jr & Tenney, S.M. (1970). *Respir. Physiol.* **10**, 384–395.
Benchetrit, G., Shea, S.A., Pham Dinh, T., Bodocco, S., Baconnier, P. & Guz, A. (1989). *Respir. Physiol.* **75**, 199–210.
Drorbaugh, J.E. & Fenn, W.O. (1955). *Paediatrics* **16**, 81–87.
Galivan, G.J. & McDonell, W.N. (1988). *Can. J. Vet. Res.* **52**, 293–298.
Tobin, M.J., Chadha, T.S., Jenouri, G., Birch, S.J., Gareroglu, H.B. & Sackner, M.A. (1983). *Chest* **84** (2), 202–205.

Pulmonary receptors in the spontaneously breathing anaesthetized rat

L. Pirie and A. Davies

Department of Physiology, University of Edinburgh, Edinburgh EH8 9AG

It has been suggested that the rate of adaption of pulmonary stretch receptors should be greater in species with high respiratory rates to maintain effective reflex control of breathing (Bartlett & St John, 1979). Our alternative suggestion is that a greater proportion of rapidly adapting receptors exists in more rapidly breathing species. This increases the overall rate of adaption. To test this hypothesis the activity of seventy-three pulmonary mechanoreceptors with afferent fibres in the left vagus nerve was studied in sixteen spontaneously breathing anaesthetized (6 ml kg^{-1} 25% Urethane, I.P.) rats, during eupnoea and sustained inflation of the lungs. Fifty-one receptors discharged mainly in inspiration, and were slowly adapting (PSRs). Twelve discharged exclusively during early inspiration. Thirty discharged throughout inspiration and in early expiration. Nine discharged throughout inspiration and expiration.

Twenty-two rapidly adapting receptors (RARs) were spontaneously active during eupnoea (peak frequency 87·52 ± 6·10 Hz, mean frequency 28·77 ± 1·25 Hz), discharged almost exclusively during expiration (2·59 ± 0·23 impulses in inspiration, 13·34 ± 0·55 impulses in expiration, 110 breaths) and, by definition, totally adapted in 0·25 s.

The patterns and total numbers of discharges for RARs were remarkably constant from breath to breath for individual receptors (e.g. 12·4 ± 0·4 impulses in five consecutive expirations).

The discharge frequencies of these receptors were comparable with those of other species (Widdicombe, 1954). The abundance of RARs was greater than in larger species (Roumy & Leitner, 1980).

The high proportion of RARs may represent an evolutionary advantage in neural control of the high frequency breathing of small mammals.

(Figures are means and standard errors of the mean).

Supported by the Norman Salvesen Emphysema Research Trust.

REFERENCES

Bartlett, D. Jr & St John, W.M. (1979). *Resp. Physiol.* **37**, 303–312.
Roumy, M. & Leitner, L.M. (1980). *J. Physiol.* **76**, 67–70.
Widdicombe, J.G. (1954). *J. Physiol.* **123**, 71–104.

Low frequency heart rate components in sick and healthy human neonates

G.C. Halley, J.H. Dripps, H. Janssens and N. McIntosh

Department of Child Life and Health, University of Edinburgh, 20 Sylvan Place, Edinburgh EH9 1UW

In the adult population it is well recognized that oscillations in the heart rate at 0·05 and 0·1 Hz are present and represent thermoregulatory and blood pressure control mechanisms respectively. The objective of this observational study was to examine low frequency heart rate and respiratory rate oscillations in healthy term infants and compare these results with a group of sick infants at term.

The effects of gamolenic acid on adult atopic eczema and premenstrual exacerbation of eczema

FRANCES HUMPHREYS, JULIAN A. SYMONS, HELEN K. BROWN,
GORDON W. DUFF, JOHN A.A. HUNTER

A double blind, parallel, placebo controlled study was performed to assess the effects of four months treatment with gamolenic acid in evening primrose oil on adult atopic eczema. Fifty-eight subjects entered the study and results were analysed for fifty-two of these. Subjects were divided into three groups, women with, and women without, a reported premenstrual exacerbation of their eczema and men. Mean results for the three groups combined showed a significant effect of evening primrose oil on erythema and surface damage when compared with placebo. No significant effect on the mean clinical score for lichenification was found. When maximum severity was examined there was a significant effect of evening primrose oil on erythema after four months treatment and on lichenification 2 months post-treatment but no effect on surface damage. Serum soluble interleukin 2 receptor levels fell to a greater extent with evening primrose oil than with placebo but this was not statistically significant. Women who reported a premenstrual flare of their eczema showed a greater improvement with GLA compared with placebo than the other 2 groups. Adjunctive treatment with gamolenic acid in evening primrose oil should be considered in patients with chronic atopic eczema.

n the 1920's it was first noted that rats deprived of dietary fatty acids failed to thrive and developed skin signs similar to human atopic eczema [1]. It was found that oral admistration of the unsaturated fatty acid, linoleic acid, the prersor of gamolenic acid, reversed these changes and the rm essential fatty acid (EFA) was coined [2]. Furthermore, it as discovered that topical administration of linoleic acid to A-deficient rats restored the characteristically abnormal rrier function of the epidermis to normal [3]. Similar skin anges were also noted in human infants whose diet lacked sential fatty acids [4]. These changes were also reported in tients suffering from fat malabsorption requiring parenteral eding and the consequent skin signs and increased transedermal water loss were reversed by the application of sunflower oil, which is rich in linoleic acid [5]. This led to the supposition that dietary supplementation of some essential fatty acids might have a therapeutic role in eczema. Further work revealed that patients with atopic eczema have elevated plasma levels of the essential fatty acid, linoleic acid, but have reduced proportions of its metabolites suggesting a block in the conversion pathway [6]. However, interpretation of these findings is hindered by the fact that these substances can theoretically be precursors of anti-inflammatory as well as pro-inflammatory molecules via the cyclo-oxygenase and lipoxygenase pathways.

Oil derived from the seed of the evening primrose plant is rich in two essential fatty acids, gamolenic acid and linoleic acid, and this oil has been used in the treatment (usually adjunctive) of atopic eczema in adults and in children. In the published clinical trials to date, in which evening primrose oil has been assessed in adult atopic eczema, conflicting results have been obtained. Four studies have shown some beneficial effects [7-10] and two larger studies have concluded that there was no significant therapeutic benefit [11, 12]. In addition a meta-analysis of some previously published and some unpublished data was performed which showed a significant therapeutic effect of fatty acid supplementation [13]. This meta-analysis has been the subject of some criticism [14] and it has been suggested that differences in baseline severity between the patient groups may have accounted for an apparent effect of evening primrose oil, patients with more severe eczema having more "potential for improvement".

It has been previously noted that up to 30% of women with chronic atopic eczema report a premenstrual worsening of their symptoms [15]. In addition to its putative effects on atopic eczema, evening primrose oil has been found to have a beneficial effect in pre-menstrual mastalgia [16] and the premenstrual syndrome [17]. Furthermore there are many documented interactions between the hormonal system and arachidonic acid and its precursors.

For these reasons we designed a study to examine the effects of evening primrose oil in adult atopic eczema with particular emphasis on a group of patients who reported a premenstrual exacerbation of their symptoms.

In order to do this, severity of eczema was measured clinically by an observer, subjective assessments of severity of eczema were recorded by the patients and serum soluble interleukin 2 receptor (sIL-2R) levels were measured. These have previously been shown to be significantly raised in patients with atopic eczema compared with a control population and to correlate with disease activity [18].

The primary aim of the study was to determine whether 4 months treatment with gamolenic acid in evening primrose oil was of benefit in the treatment of adult atopic eczema. The secondary aim was to detect if the effect (if any) of fatty acid supplementation differed in patients reporting a pre-menstrual flare of their eczema.

Materials and methods

Fifty-eight adults with moderately severe atopic eczema took part in the study. All satisfied the diagnostic criteria of Hanifin and Rajka [19]. The duration of the study was thirty weeks in total comprising a four weeks run-in period, sixteen weeks treatment period and a further eight weeks evaluation after stopping treatment. Female patients were assessed on six occasions, a midcycle and a premenstrual visit being made (usually 14 days apart) before treatment, during the last month of treatment and in the last month of the study. Male patients were assessed on five occasions, two visits a fortnight apart being made prior to treatment, in the last month of treatment and one visit being made two months after termination of treatment. At each visit a clinical scoring system was used which scored severity of three parameters, erythema, surface damage and lichenification on a scale of 0-3. The anterior and posterior surfaces of the body were each divided into ten regions and the percentage surface area affected by each of the three clinical signs of eczema was assessed. The severity score for each parameter was then multiplied by the percentage area affected and the scores for each of the twenty areas added up to give a body total score for each parameter [20]. All containers of topical corticosteroid in use or finished since the previous visit were weighed and new weighed tubes were issued in order to estimate topical corticosteroid usage. In addition to this serum was collected at each visit and stored for subsequent measurement of sIL-2R levels. Serum sIL-2R was measured by ELISA as previously described [21] (T Cell Sciences Inc.; Laboratory Impex Ltd, U.K.). All samples were coded and read blind in the assay.

All subjects kept diaries which were filled in daily by the female participants and weekly by the males. In the diaries severity of itch, dryness, scaling, redness and overall assessment of eczema were recorded on 100 mm visual analogue scales by all participants and female subjects also assessed severity of six premenstrual symptoms (breast pain, bloatedness, depression, irritability, headache and clumsiness).

All patients continued during the study with their usual treatment in the form of topical corticosteroids, emollients and any systemic treatment. Treatment with evening primrose oil and placebo was randomly allocated (by computerised block allocation) on a double blind basis. The active treatment given was gamolenic acid in evening primrose oil (GLA) 500 mg oil plus 10 mg vitamin E in gelatin capsules (Epogam), twelve capsules daily. The placebo used was liquid paraffin 500mg with 10 mg vitamin E in identical capsules. Vitamin E is included in these preparations as an anti-oxidant.

All patients were included in the main analysis. Clinical scores were analysed before treatment (stage 1), in the last month of treatment (stage 2) and two months after treatment (stage 3). As patients visited twice during each stage, mean values of the clinical scores were used and maximum severity over the two visits was also examined. All results were analysed blind using analysis of covariance. In this standard method of statistical analysis baseline measurements were fitted as the covariate for each score. This reduces variations caused by random differences between groups at the start of the study, and also corrects for any influence of baseline score on final score [22]. Using analysis of covariance results in the production of "adjusted means" for comparison between the active treatment and the placebo groups.

In addition, the women were studied to examine whether a treatment by cycle interaction existed i.e. whether women reporting a premenstrual exacerbation of their eczema (Group A) reacted differently to treatment from women who did not (Group B), at different stages of the menstrual cycle; and whether the women as a whole reacted differently at different stages of the menstrual cycle.

SIL-2R levels and weight of topical corticosteroid used were similarly examined by analysis of covariance.

The data from the patients' self-assessments were analysed using Fortran programmes to calculate mean values for each symptom at different stages of treatment. In addition, the women's self-assessment data were analysed to determine any change in symptoms at different stages of the menstrual cycle.

The data were also examined to determine whether there were any interactions between treatment and group, age or oral contraception.

This research was approved by the Lothian Research Ethics Committee (Medicine/Clinical Oncology).

Results

Fifty-eight patients entered the study, twenty-six were women who reported definite consistent premenstrual exacerbation of their eczema (Group A), seventeen were women without exacerbation (Group B) and fifteen were men (Group C).

Three patients attended only for the initial explanatory visit (all placebo). Three subjects defaulted within the first month of treatment (two active, one placebo). These six subjects were not included in the analysis because no post-treatment data were available. All other subjects were included and if any defaulted later in the study the previous attendance was used as the endpoint. Results from fifty-two patients were analysed, twenty-six in Group A, twelve in Group B and fifteen in Group C.

Demographic data are shown in *Table I* and show an even distribution of subjects between the GLA and placebo groups in terms of sex, age and duration of eczema.

The clinical scores for erythema, surface damage and lichenification were found to follow a normal distribution when logged and therefore log scores were used in each analysis.

Table II shows the mean scores for all patients for the three clinical parameters measured before and after 4 months treatment with GLA and placebo and 2 months post-treatment. The GLA group has higher pretreatment clinical scores for erythema and surface damage. These differences are not statistically significant for erythema. A significant difference was, however, found for surface damage (p = 0.04). The mean scores for erythema and surface damage fell during active treatment but not with placebo. *Figure 1* shows the individual changes in erythema with GLA and placebo. Twenty-three out of 27 patients taking active treatment showed an improvement in their clinical score for erythema by the end of the treatment period compared with 11/23 in the placebo group. The results for surface damage *(Fig. 2)* were very similar, 12/23 in the placebo group showing an improvement in clinical score, compared with 23/27 in the GLA group. *Table III* shows the results of the analysis of covariance of the clinical scores. The adjusted means after sixteen weeks GLA and placebo and the p-values are shown. Highly significant differences in erythema and surface change were detected for GLA compared with placebo after four months treatment and also post-treatment. No significant difference was found for lichenification.

When maximum severity was analysed using an ordered logistic regression model a significant treatment effect was found for erythema at sixteen weeks (p = 0.05) and for liche-

Table I. ***Demographic data relating to subjects***

	n	Age range	Age at onset	Duration
GLA	28	18-49	1 m-30 y	3-47 y
Placebo	24	16-64	3 m-21 y	2-47 y

Table II. ***Clinical scores for all patients for erythema, surface damage and lichenification before, after 4 months treatment and 2 months post-treatment (means ± s.e.m.), GLA n = 27, placebo n = 25***

		Before	16 weeks	Post-
Erythema	GLA	397 ± 74	268 ± 63	339 ± 156
	Placebo	246 ± 45	240 ± 46	318 ± 53
Surface damage	GLA	411 ± 70	277 ± 71	313 ± 145
	Placebo	237 ± 40	235 ± 49	285 ± 47
Lichenification	GLA	99 ± 23	136 ± 75	65 ± 19
	Placebo	122 ± 29	97 ± 35	178 ± 67

Table III. ***Analysis of covariance. The adjusted means after treatment with gamolenic acid and placebo, 95% confidence intervals and p-values, GLA n = 27, placebo n = 25***

		Erythema	Surface damage	Lichenification
GLA	mean	111	102	24
	95% CI	(85-145)	(79-138)	(15-51)
Placebo	mean	179	184	17
	95% CI	(135-240)	(154-245)	(9-35)
p-values		0.017	0.007	0.2

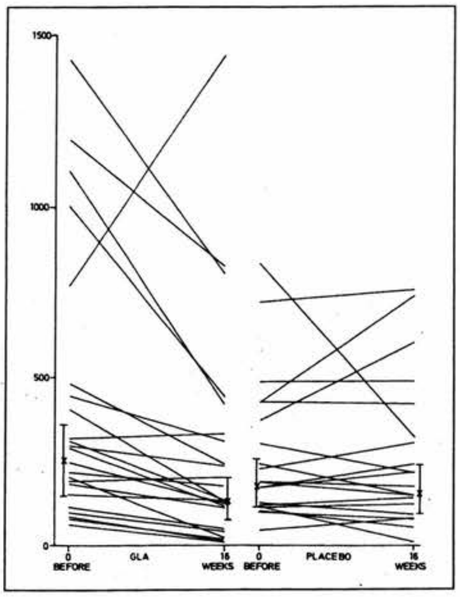


Figure 1. ***Individual changes in erythema with gamolenic acid (n = 27) and placebo (n = 25).*** *Means and 95% confidence intervals shown.*

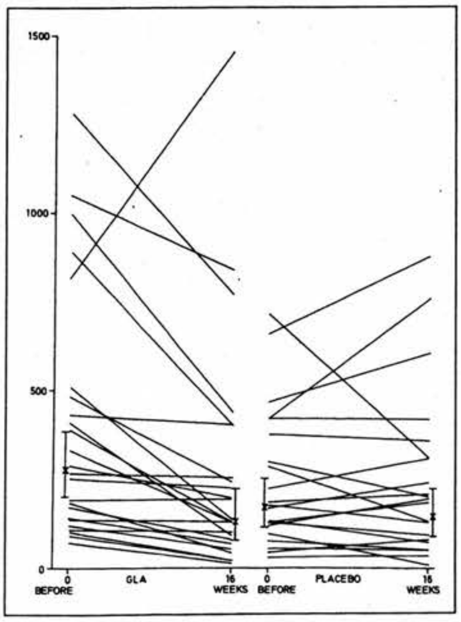


Figure 2. ***Individual changes in surface damage with gamolenic acid (n = 27) and placebo (n = 25).*** *Means and 95% confidence intervals shown.*

nification two months post-treatment (p = 0.04) where GLA patients had less severe scores.

Investigation of treatment by cycle interactions in female patients showed no conclusive evidence that the two groups of women reacted differently to treatment at different stages of their menstrual cycle. There was also no evidence that women as a whole reacted differently to treatments at different stages of the menstrual cycle.

To further examine the effect of active treatment and placebo in the different patient groups, reduction in clinical score as a percentage of original clinical score was calculated for each individual and the mean ± s.e. mean found. These results for erythema *(Fig. 3)* and surface damage are illustrated *(Fig. 4)*. No statistically significant differences were found but patients in Group A (women who reported a premenstrual flare) showed the greatest difference in response between GLA and placebo.

The sIL-2R measurements were found to follow a normal distribution when logged and log scores were also used in their analysis. *Table IV* shows the mean sIL-2R levels before treatment, after 16 weeks treatment and 2 months later. The fall in sIL-2R levels was significant for both GLA and placebo after four months' treatment and for GLA 2 months post-treatment by the paired t-test. This fall was greater for the GLA group but there was no statistically significant difference between GLA and placebo when compared by analysis of covariance. After treatment it can be seen that sIL-2R levels remained reduced in the GLA group but rose in the placebo group.

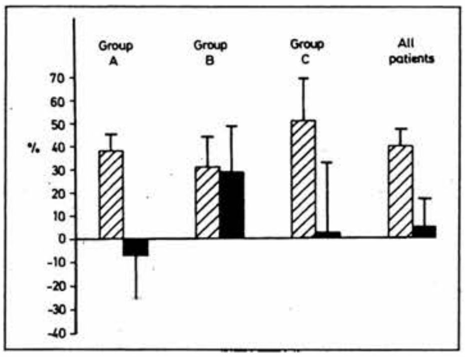


Figure 3. ***Mean reduction in erythema at the end of treatment as a percentage of the individual's original score for the three patient groups and all patients combined.*** *(▨ = gamolenic acid, ■ = placebo).*

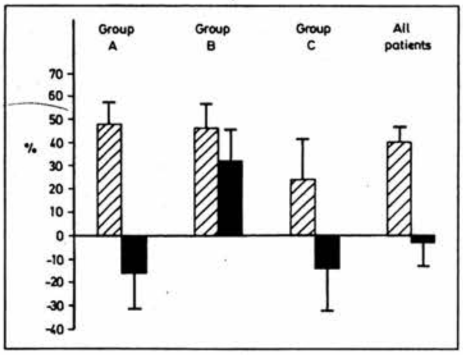


Figure 4. ***Mean reduction in surface damage at the end of treatment as a percentage of the individual's original score for the three patient groups and all patients combined.*** *(▨ = gamolenic acid, ■ = placebo).*

There was no significant change in the amount of topical corticosteroid used in either the GLA or the placebo group. Prior to the treatment period the active group were applying a mean ± s.e. of 5.0 ± 1.1 g potent topical steroid (betamethasone valerate 0.025% or alternative topical steroid of equivalent potency) daily and the placebo group were using 3.5 ± 1.0 g compared with 5.1 ± 1.3 g and 3.1 ± 0.8 g respectively at the end of the 4 month treatment period.

The pre-treatment data were analysed to determine whether any factors were related to the three main clinical scores. No factor was found to be statistically significantly related to any score when all patients were looked at together. When women only were examined it was found that scores for redness and surface damage were higher at the midcycle than the premenstrual visit.

We also examined the hypothesis that patients with the worst pre-treatment clinical scores would show the greatest improvement during the treatment period [14]. The relationship between pre- and post-treatment scores did not differ between the active and the placebo group. There was a strong trend for those with the least severe erythema scores to improve most when active treatment was compared with placebo (p = 0.07).

Results from the patients' diaries are summarised in *Table V*. This table shows the means and s.e. ms for all patients prior to treatment, from 8 weeks to the end of treatment and in the post-treatment period. There were no statistically significant changes found between the GLA and placebo groups by ana-

Table IV. ***Mean sIL-2R levels (U/ml) before and after 4 months' treatment, and two months post treatment, the effect of gamolenic acid and placebo (paired t-test), GLA n = 27, placebo n = 25***

		Baseline	**16 weeks**	**24 weeks**
GLA	mean	897	727	742
	95% CI	750-1075	608-871	567-925
	p-value		0.002	0.007
Placebo	mean	705	607	706
	95% CI	620-804	507-699	567-846
	p-value		0.018	0.665

Table V. ***Diary data all patients,* (n = 52) mean ± s.e.m.**

	Pre-treatment		**End-treatment**		**Post-treatment**	
	GLA	**Placebo**	**GLA**	**Placebo**	**GLA**	**Placebo**
Itch	42 ± 3	39 ± 3	36 ± 3	37 ± 3	43 ± 4	43 ± 4
Dryness	40 ± 3	35 ± 3	35 ± 3	35 ± 3	42 ± 4	39 ± 4
Scaling	33 ± 3	28 ± 4	29 ± 4	28 ± 3	35 ± 4	31 ± 5
Erythema	36 ± 4	34 ± 3	33 ± 4	37 ± 4	37 ± 5	38 ± 4
Overall	42 ± 3	37 ± 3	37 ± 3	37 ± 3	42 ± 4	40 ± 4

lysis of covariance. A cyclical variation in diary symptoms, apparently related to the menstrual cycle was noted in 9/23 women who reported a premenstrual flare of eczema and in 5/16 who did not. However, no statistically significant variation in diary symptoms was found at different times of the menstrual cycle in the pre-treatment month in Group A or Group B. Scores for redness were recorded both by the clinician and the patient, and a reasonable correlation between these measurements was found (Spearman rank coefficients; pre-treatment 0.631, end treatment 0.592).

Two patients in the GLA group developed diarrhoea during the study, one of whom also developed colicky abdominal pain. One patient in the placebo group reported similar symptoms. These symptoms were lessened by increasing the frequency and decreasing the dose of capsules taken at any one time. One patient in the GLA group experienced worsening of her atopic eczema and an urticarial rash one week after starting treatment. She had experienced a similar reaction on a previous occasion when taking evening primrose oil. Treatment was stopped. Her results at the time of withdrawal from the study were included in the analysis.

Discussion

These data show a beneficial effect of 4 months' treatment with GLA on erythema and surface damage in adult atopic eczema. A significant effect on lichenification was found only when maximum severity was examined and then only two months post-treatment. The sIL-2R levels fell during the treatment period in both the active and the placebo groups and no statistically significant difference between the groups was found.

It is interesting that our study showed a positive result when most previously reported studies have shown a small beneficial effect of GLA, which was not necessarily statistically significant, or no significant benefit. The women in our study who reported a premenstrual flare of their symptoms (Group A) showed a greater, though not statistically significantly different, response to GLA when compared with placebo than the other two groups *(Fig. 3)* and it is possible that our choice of patients has resulted in these findings.

The patients in the GLA group had mean pre-treatment scores greater than those of the placebo group. As it has been suggested that there may be a greater chance of detecting an improvement when this is the case [14], we adopted analysis of covariance in our treatment of the data, using the pretreatment scores as the co-variate in order to minimise this effect. Furthermore, our analysis of the data has shown that those patients with the lower scores pre-treatment improve most, and not the converse.

There are many putative mechanisms to explain a connection between essential fatty acids and the pre-menstrual syndrome in which low plasma levels of essential fatty acid metabolites have been found [23]. Neither the clinical scoring system we used nor the patients' self-assessment data confirmed a premenstrual exacerbation of eczema in Group A patients in our study in the pre-treatment assessments. The mean clinical scores were greatest at day ten and the best response to treatment was seen at this time and not at day twenty-three of the menstrual cycle. It is possible therefore that perception of symptoms plays a role in reporting of premenstrual exacerbation of eczema. It is also possible that the one month pretreatment phase was not adequate to demonstrate any premenstrual exacerbation. The Group A patients showed a greater difference in effect of GLA when compared with placebo, and a smaller placebo response. Although this did not reach statistical significance, we believe this should be examined further.

Although no statistically significant difference was demonstrated between the active and the placebo groups, it was interesting to note a greater and more sustained fall in sIL-2R levels in the patients taking GLA. Such a fall could reflect reduced inflammation consequent upon clinical improvement in their eczema. However, there is evidence that essential fatty acids have direct effects on production of interleukin 1, tumour necrosis factor and interleukin 2 *in vitro*, and this is an alternative explanation [24].

We did not find any fall in mean topical corticosteroid usage in either group during the course of the study. However, we noted that only the minority of patients were entirely reliable in returning used containers.

Although we found a clear improvement in clinical scores statistically, for most patients this meant a modest improvement in the severity of their eczema clinically. However, the incidence of adverse events was very low. We feel, therefore, that in the context of a chronic skin condition, treatment with GLA should be considered and may be particularly effective in women reporting a premenstrual flare of their symptoms. ■

Acknowledgements

We thank Dr D. Kemmett for providing many of the patients and Dr R. Elton, Department of Medical Statistics, University of Edinburgh for further statistical advice. Dr F. Humphreys was supported by a grant from Scotia Pharmaceuticals.

Synopsis

A double blind, placebo controlled study of the effect of four months' treatment with gamolenic acid in evening primrose oil on adult atopic eczema was performed. Fifty-eight patients were studied including 26 women who reported premenstrual exacerbation of symptoms. A significant effect of treatment was found on erythema and surface damage, but not on serum soluble interleukin 2 receptor levels.

F. Humphreys: University Department of Dermatology, Level 4, Lauriston Building, The Royal Infirmary, Edinburgh EH3 9YW, UK.
J.A.A. Hunter: Department of Dermatology, University of Edinburgh.
J.A. Symons, G.W. Duff: Section of Molecular Medicine, Department of Medicine, University of Sheffield.
H.K. Brown: Unit of Medical Statistics, Department of Public Health, University of Edinburgh, United Kingdom.

Reprints: F. Humphreys.

Key words: atopic eczema, essential fatty acids, interleukin 2 receptors, the menstrual cycle, double blind placebo controlled study.

References

1. Burr GO, Burr MM. A new deficiency disease produced by the rigid exclusion of fat from the diet. *J Biol Chem* 1929; 82: 345-67.

2. Burr GO, Burr MM. On the nature and role of the fatty acids essential in nutrition. *J Biol Chem* 1930; 86: 587-621.

3. Hartop PJ, Prottey C. Changes in transepidermal water loss and the composition of epidermal lecithin after application of pure fatty acid triglycerides to the skin of essential fatty acid deficient rats. *Br J Dermatol* 1976; 95: 255-64.

4. Hansen AE, Wiese HF, Boelsche AN, *et al.* Role of linoleic acid in infant nutrition. Clinical and chemical study of 428 infants fed on milk mixtures varying in kind and amount of fat. *Pediatrics* 1963; 31: 171-92.

5. Prottey C, Hartop PJ, Press M. Correction of the cutaneous manifestations of essential fatty acid deficiency in man by application of sunflower-seed oil to the skin. *J Invest Dermatol* 1975; 64: 228-34.

6. Manku MS, Horrobin DF, Morse NL, *et al.* Essential fatty acids in the plasma phospholipids of patients with atopic eczema. *Br J Dermatol* 1984; 110: 643-8.

7. Lovell CR, Burton JL, Horrobin DF. Treatment of atopic eczema with evening primrose oil. *Lancet* 1981; 1: 278.

8. Wright S, Burton JL. Oral evening primrose seed oil improves atopic eczema. *Lancet* 1982; 2: 1120-2.

9. Schalin-Karrila M, Mattila L, Jansen CT, Uottila P. Evening primrose oil in the treatment of atopic eczema: effect on clinical status, plasma phospholipid fatty acids and circulating blood prostaglandins. *Br J Dermatol* 1987; 117: 11-9.

10. Sugai T. Clinical evaluation of oral evening primrose oil (Efamol) in atopic patients. *Skin Res (Tokyo)* 1987; 29: 330-8.

11. Bamford JTM, Gibson RW, Reiner CM. Atopic eczema unresponsive to evening primrose oil (linoleic and gamma-linolenic acids). *J Am Acad Dermatol* 1985; 13: 959-65.

12. Berth-Jones J, Graham-Brown RAC. Placebo-controlled trial of essential fatty acid supplementation in atopic dermatitis. *Lancet* 1993; 341: 1557-60.

13. Morse PF, Horrobin DF, Manku MS, Stewart JCM. Meta-analysis of placebo-controlled studies of the efficacy of Epogam in the treatment of atopic eczema: relationship between plasma essential fatty acid changes and clinical response. *Br J Dermatol* 1989; 121: 75-90.

14. Sharpe GR, Farr PM. Evening primrose oil and eczema. *Lancet* 1990; 335: 667-8.

15. Kemmett DK, Tidman MJ. The influence of the menstrual cycle and pregnancy on atopic dermatitis. *Br J Dermatol* 1991; 125: 59-61.

16. Pye JK, Mansel RE, Hughes LE. Clinical experience of drug treatments for mastalgia. *Lancet* 1985; 330: 373-7.

17. Puolakka J, Makarainen L, Viinikka L, Ylikorkala O. Biochemical and clinical effects of treating the premenstrual syndrome with prostaglandin synthesis precursors. *J Reprod Med* 1985; 30: 149-53.

18. Colver GB, Symons JA, Duff GW. Soluble interleukin 2 receptor in atopic eczema. *Br Med J* 1989; 298: 1426-8.

19. Hanifin JM, Rajka G. Diagnostic features of atopic dermatitis. *Acta Dermatovener* 1980; suppl 92: 44-7.

20. Heddle RJ, Soothill JF, Bulpitt CJ, Atherton DJ. Combined oral and nasal beclomethasone dipropionate in children with atopic eczema: a randomised controlled trial. *Br Med J* 1984; 289: 651-4.

21. Symons JA, Wood NC, DiGiovine FS, Duff GW. Soluble IL-2 receptor in rheumatoid arthritis: correlation with disease activity, IL-1 and IL-2 inhibition. *J Immunol* 1988; 141: 2612-8.

22. Armitage P, Berry G. *Statistical methods in medical research.* Chapter 9, 2nd edition. Oxford, UK: Blackwell Scientific Publications 1987: 282-95.

23. Horrobin DF, Manku MS. Premenstrual syndrome and premenstrual breast pain (cyclical mastalgia): disorders of essential fatty acid metabolism. *Prostaglandins, leukotrienes and essential fatty acids: Reviews* 1989; 37: 255-61.

24. Zurier RB. Fatty acids, inflammation and immune responses. *Prostaglandins, leukotrienes and essential fatty acids* 1993; 48: 57-62.

STATISTICS IN MEDICINE, VOL. 13, 1601–1617 (1994)

THE APPLICATION OF REML IN CLINICAL TRIALS

H. K. BROWN

Medical Statistics Unit, Department of Public Health Sciences, University of Edinburgh, EH8 9AG, U.K.

AND

R. A. KEMPTON

Scottish Agricultural Statistics Service, King's Buildings, (JCMB), Edinburgh EH9 3JZ, U.K.

SUMMARY

Residual maximum likelihood (REML) is a technique for estimating variance components in multi-classified data. In contrast to analysis of variance it can be routinely applied to unbalanced data and avoids some of the problems of biased variance estimates found with standard maximum likelihood estimation. The full REML method is of particular value for the analysis of unbalanced clinical trials as it allows recovery of all the available information on treatment effects which can lead to significant improvements in their precision. The use of REML has until recently been limited by heavy computational requirements and lack of readily available software. This is no longer such a restriction, however, as REML procedures are now available in several widely-used statistical packages, including BMDP, Genstat and SAS. This paper describes the REML technique and discusses its application to three common types of clinical trial: crossover, repeated measures and multicentre.

1. INTRODUCTION

In a clinical trial, variation in response may arise from several sources associated with different strata of the experiment. For example, in a crossover or repeated measures trial, there will be a component of variance for responses on the same patient and a component for responses of different patients. Similarly, in a multicentre trial, separate components of variance can be identified among patients from the same centre and among the different centres. The stratum variance components can be used to increase the precision of treatment estimates by allowing information to be combined efficiently across strata using the method of generalized least squares. Knowledge of the stratum variances may also help to improve treatment design in subsequent experiments.

Fisher,[1] in his book *Statistical Methods for Research Workers*, introduced the analysis of variance (ANOVA) to a wide audience and outlined the basic method for estimating components of variance by equating the mean squares from an ANOVA table to their expected values. Yates[2] and Henderson[3] showed how Fisher's technique could be extended to unbalanced (that is, incomplete) data, but there is then no unique way of choosing the appropriate mean squares and the ANOVA estimators are not fully efficient. Hartley and Rao[4] proposed using maximum likelihood (ML) to provide unique estimators of variance components. However, the ML estimates are generally biased downwards as the method assumes that fixed effects are already

CCC 0277–6715/94/161601–17

Received March 1993
Revised October 1993

known, rather than also being estimated from the data. Indeed, in the balanced case, ML variance estimators are generally smaller than the unbiased estimators from ANOVA.

Residual maximum likelihood (REML) overcomes the problem of bias by automatically adjusting for the degrees of freedom corresponding to estimated treatment effects, as is done by ANOVA for balanced data. Indeed, for balanced data, REML and ANOVA estimators are identical. Also, when only one variance stratum is defined and all other effects are taken as fixed, REML is equivalent to simple least squares analysis. The general technique for REML was developed by Patterson and Thompson.[5,6] The underlying mixed model and estimating equations are given in the Appendix. Searle *et al.*[7] provide a more detailed theoretical description of REML and other variance estimation methods.

REML has been applied widely in agricultural research. It was introduced by Patterson and Thompson[5] in the context of incomplete block designs where the precision of treatment estimates may be improved by including between-block information. The method has been used extensively in animal breeding to estimate heritabilities and predict genetic gain from breeding programmes.[8,9] Talbot[10] used REML to estimate variance components for variety trialling systems carried out across several centres and years for different crops and was thus able to compare their general precision and effectiveness. A wider review of applications is provided by Robinson.[11]

REML can also make an important contribution to the analysis of clinical trials. These are frequently unbalanced either by design or due to events such as early study termination, missing records, patient withdrawals, exclusions or incorrect randomizations. It is ethically important that as few patients as possible are used in a trial and this requires that trials are efficiently designed and analysed. However, in many analyses, information is lost by using least squares estimation which ignores information from higher strata, or by deliberately discarding data to ensure the trial data are balanced for treatments. Several authors[12–14] have described methods for using this information to improve the precision of special types of clinical trial. REML is a general method that can be used for any multi-classified dataset and ensures an efficient analysis regardless of the level of data imbalance.

REML techniques are computationally relatively costly and their restricted use could until recently be explained by lack of readily available software. However, these drawbacks have been largely overcome by the introduction of REML procedures in widely used general statistical packages, including SAS,[15] BMDP[16] and Genstat.[17] There is now a need to build up a body of experience of applying the REML method so that it may be used confidently on a routine basis in clinical trials.

As part of this process, our paper describes the application of REML in three different kinds of clinical trial – crossover, repeated measures and multicentre – and outlines REML facilities currently available in statistical computer packages.

2. CROSSOVER TRIALS

Crossover designs are used extensively in clinical trials to compare the efficiency of treatments applied sequentially to patients. These designs have the advantage that treatment differences can be compared against a background of within-patient variation, rather than the typically larger between-patient variation.

When all patients receive all treatments equally, a conventional least squares analysis based on within-patient comparisons is fully efficient for the simple additive treatment effects model. However, crossover trials are often not fully balanced, either by design, for example when the number of periods is less than the number of treatments, or, more commonly, due to missing data.

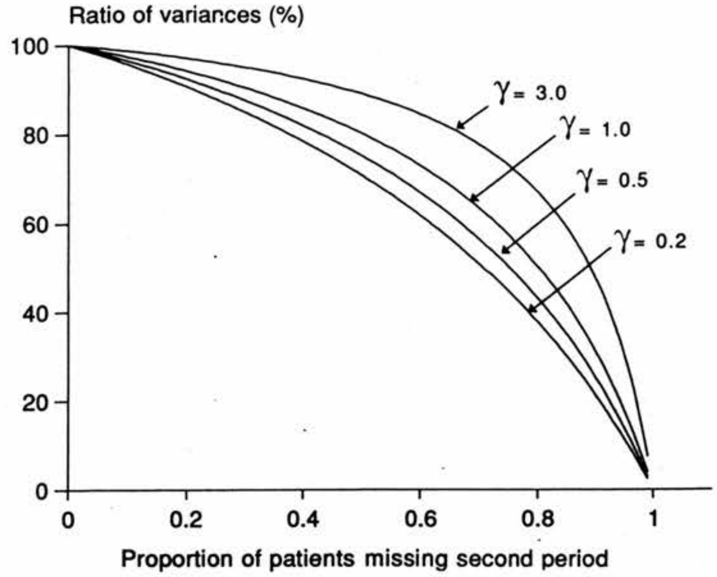


Figure 1. Variance of difference in treatment effects with full recovery of between-patient information as a proportion of the variance with no recovery for a two-period cross-over design with specified proportion of patients missing in the second period. γ is the ratio of between- to within-patient variance components

Additional information on treatment comparisons is then available between patients and may be used to increase the precision of estimates.

The amount of extra information depends on the level of treatment imbalance and the relative sizes of between- and within-patient variances. We illustrate this with a simple example. Consider a two-period crossover trial for two treatments, A and B, with N patients in each treatment group, and assume that, for a proportion p of patients in each group, only the response for the first period is recorded. The variance in response within patients is σ^2, while the variance between patients is $(1 + \gamma)\sigma^2$. Then the variance of the usual estimate for difference in treatment means for the $2N(1 - p)$ within-patient comparisons is $\sigma_w^2 = \sigma^2/N(1 - p)$, while the corresponding variance from between-patient comparisons for period 1 responses of the remaining $2Np$ patients is $\sigma_b^2 = 2\sigma^2(1 + \gamma)/Np$. The variance of the combined estimate of treatment difference from pooling within- and between-patient information is then $2\sigma^2(1 + \gamma)/N[p + 2(1 + \gamma)(1 - p)]$. So using the between-patient information reduces the variance of treatment differences to a proportion $1 - p/[p + 2(1 + \gamma)(1 - p)]$ of the within-patient variance. This is plotted in Figure 1 for different values of p and γ. These theoretical results are derived with the assumption that γ is known. In practice γ is estimated and there may be a slight loss of information for the combined estimate in small samples (see Section 7).

Yates[2] described a method for recovering between-block information for balanced incomplete block designs, but the benefits of incorporating this information in crossover studies is only now becoming widely recognized.

Gough[18] used REML with random patient effects to recover between-patient information in a two-treatment, three-period crossover study,[19] and showed that this led to a reduction of 20 per cent in the variance of treatment estimates. Chi[14] outlined an analysis of incomplete block trials with recovery of information and applied it to three crossover trials:[20] this led to reductions of between 3 and 8 per cent in variances of treatment estimates. Both authors made allowance for

Table I. Variances of estimates of differences of treatment and carryover effects for a three-treatment, two-period design (in multiples of $2\sigma^2/r$). γ is the ratio of the between-patient to within-patient variance components

	Within-patient analysis	Between-patient analysis	Combined analysis
Treatment effects			
Omitting carryover	$\frac{1}{3}$	$1 + 2\gamma$	$\frac{1 + 2\gamma}{2(2 + 3\gamma)}$
Including carryover	$\frac{4}{3}$	$\frac{4}{3}(1 + 2\gamma)$	$\frac{2(1 + 2\gamma)(1 + \gamma)}{7 + 14\gamma + 3\gamma^2}$
Carryover effects	4	$\frac{4}{3}(1 + 2\gamma)$	$\frac{2(1 + 2\gamma)(2 + 3\gamma)}{7 + 14\gamma + 3\gamma^2}$

treatment effects being carried over to the next period in their analyses. Including carryover effects in the response model will generally reduce the information on treatment effects from the within-patient comparisons and consequently increase the importance of using the additional between-patient information, particularly for two-period crossover trials.[21]

Here we consider a simple case of the three-treatment (A, B and C), two-period design with r replicates of the six sequences AB, BA, AC, CA, BC and CB. The variances of treatment differences for within- and between-patient comparisons and a combined analysis are expressed as functions of γ in Table I, for models both with and without carryover effects. In the absence of carryover effects, the within-patient analysis will generally have high efficiency as the variance of treatment estimates is never more than 4/3 of the variance for the full analysis. (This upper limit occurs when $\gamma = 0$ and corresponds to the reciprocal of the efficiency of the design.) The inclusion of carryover effects, however, leads to a four-fold increase in treatment variance from within-patient analysis. There is however, substantial between-patient information on both treatment and carryover effects and this information may be used in a combined overall analysis to provide more precise treatment estimates. Then, for small γ, the variance of treatment estimates is increased by a factor of only $(8 + 4\gamma)/7$ when carryover effects are included in the model.

To find out what values of γ are likely to be met in practice we have analysed seven examples of crossover trials given by Jones and Kenward.[20] These cover a range of diseases and conditions and five types of design (Table II). A standard model with period, treatment and carryover effects as fixed and patient effects as random was fitted. REML estimates of γ ranged from 0·5 to 27·2 with a median of 1·5. Table II also gives, for three of these trials, the γ estimates derived by Chi[14] using an ANOVA method. This method of estimation is generally inefficient and only equivalent to REML for balanced complete or incomplete block designs.

Knowledge of the likely value of γ can help with choice of design. Suppose, for example, we wish to compare three treatments with a fixed number of patient sessions. Should we use a balanced two-period crossover trial, as above, or a parallel group trial with one treatment per patient? For the parallel group trial, there will be $4r$ patients receiving each treatment and the variance of treatment differences will be $\sigma^2(1 + \gamma)/2r$. If there are no carryover effects, the variance of treatment differences from the crossover trial, including between-patient information (Table I), will always be smaller than the parallel group trial. However, when carryover effects are present, the crossover trial is expected to be less precise when $\gamma < 1$; the potential increase in treatment precision from using within-patient comparisons is then more than offset by the aliasing with

Table II. REML estimates of ratio γ of between-patient to within-patient components of variance for seven crossover trials from Jones and Kenward.[17] Figures in brackets are ANOVA estimates for three trials derived by Chi[14]

Table[*]	Measurement (study group)	Number of patients	Design[†]	$\hat{\gamma}$
2.1	FEV[‡](bronchial asthma)	17	2 × 2	6·8 (4·0)
2.13	Plaque score (unknown)	34	2 × 2	0·5
2.19	Log (EFT)[§](women undergoing IVF)	23	2 × 2	3·7
4.3	Symptom score (Parkinson's disease)	17	2 × 2[¶]	27·2 (28·1)
4.9	Systolic BP (hypertension)	89	3 × 2	1·5 (1·5)
5.20	Systolic BP (hypertension)	23	3 × 3	1·4
5.23	LVET[‖](intermittent claudication)	14	4 × 4	0·8

[*] Table number in Jones and Kenward[17]

[†] periods × treatments

[‡] Forced expired volume in one second

[§] Visual/spatial test score

[‖] Left ventricular ejection time

[¶] four treatment sequences AA, BB, AB, BA

carryover effects. Note, however, that there are additional gains from using crossover trials if costs depend on the number of patients in the trial, not simply on the number of sessions.

2.1. A crossover trial of analgesic drugs

Mead[22] gives results of a two-period crossover trial to compare three analgesic drugs labelled A, B and C. The trial involved 43 patients in total and the numbers receiving each treatment combination were as follows: AB 7; BA 5; AC 7; CA 8; BC 8; CB 8. The effectiveness of each treatment was assessed by the numbers of hours of pain relief provided. The design did not include a washout period between treatments so there was a strong possibility of treatment carryover effects. The fitted model was

$$\begin{aligned}\text{Response} = {} & \text{Overall mean} + \text{Patient effect} + \text{Period effect} + \text{Treatment effect}\\ & + \text{Carryover effect} + \text{Random error.}\end{aligned}$$

Period, treatment and carryover effects were taken as fixed. REML analyses were carried out with patient effects specified first as fixed, then as random to allow recovery of between-patient information. The estimated variance components and treatment effects are shown in Table III, first omitting then including carryover effects. Comparison of the two analyses omitting carryover effects shows that the average standard error for treatment differences from the combined analysis with recovery of between-patient information is on average 12 per cent less than for the within-patient analysis. The comparison A–B which has the largest standard error shows greatest increase in precision, so that the combined analysis also leads to treatment estimates with a smaller range of standard errors. For this trial the between-patient component of variance is relatively small and recovery of between-patient information is clearly worthwhile.

Although the estimated carryover effects were not statistically significant, they are sufficiently large to consider including in the model. This approach is advocated by Abeyasekera and Curnow.[23] When this is done the precision of estimates of treatment effects for the within-patient analysis is dramatically reduced, with a doubling of their standard errors. By contrast, the

Table III. REML estimates of variance components and treatment effects for a crossover trial comparing three analgesic drugs[22]. (Standard errors of estimates appear in brackets.)

	Hours pain relief	
	fixed patients	random patients
Ignoring carryover		
Variance components		
patients	—	1·3 (1·89)
units within patients	10·8 (2·42)	10·7 (2·39)
Treatment effects		
A-B	3·4 (1·05)	3·5 (0·91)
A-C	2·0 (0·99)	1·9 (0·88)
B-C	– 1·4 (0·98)	– 1·6 (0·87)
Including carryover		
Variance components		
patients	—	1·3 (1·96)
units within patients	11·3 (2·60)	10·9 (2·44)
Treatment effects		
A-B	3·7 (2·01)	3·8 (0·99)
A-C	2·6 (2·14)	1·9 (0·99)
B-C	– 1·1 (2·08)	– 1·9 (0·99)
Carryover effects		
A-B	0·4 (3·43)	1·0 (1·45)
A-C	1·1 (3·66)	0·2 (1·41)
B-C	0·7 (3·72)	– 0·8 (1·46)

Note that Mead's[22] treatments estimate for the random patients model without carryover are incorrect as his values for the standard errors of treatment differences from between-patient comparisons are out by a factor of $\sqrt{2}$

combined analysis with random patient effects shows only a small increase in standard errors reflecting the improved estimation of carryover effects indicated in Table II.

For this trial the between-patients variance component is small compared to the within-patients component ($\hat{\gamma} = 0{\cdot}12$), suggesting that there is little advantage from using a crossover trial for testing these analgesics, even if carryover effects can be avoided. Indeed, the predicted average standard error for treatment comparisons for a parallel group trial with the same number of patient sessions per treatment is 0·92, compared with 0·89 and 0·99 for the crossover analysis ignoring and including carryover effects, respectively.

3. REPEATED MEASURES TRIALS

A sequence of observations are often made on each patient in clinical trials to observe the long term effect of the treatments or their variability in reponse. For example, measurements of systolic blood pressure or concentration of a test drug in the blood may be taken at intervals after treatment. Responses for successive time periods in these repeated measures trials are frequently correlated and it is important to allow for this in the statistical analysis.[20] This may not be straightforward using conventional least squares methods, particularly when the data are

Table IV. Possible covariance structures for four repeated measurements per patient

(i) General model	(ii) Autoregressive model
$V=\begin{vmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{12} & \sigma_2^2 & \theta_{23} & \theta_{24} \\ \theta_{13} & \theta_{23} & \sigma_3^2 & \theta_{34} \\ \theta_{14} & \theta_{24} & \theta_{34} & \sigma_4^2 \end{vmatrix}$	$V=\begin{vmatrix} \sigma^2 & \theta_1 & \theta_2 & \theta_3 \\ \theta_1 & \sigma^2 & \theta_1 & \theta_2 \\ \theta_2 & \theta_1 & \sigma^2 & \theta_1 \\ \theta_3 & \theta_2 & \theta_1 & \sigma^2 \end{vmatrix}$
(iii) Uniform covariance model	(iv) Uncorrelated model
$V=\begin{vmatrix} \sigma^2 & \theta & \theta & \theta \\ \theta & \sigma^2 & \theta & \theta \\ \theta & \theta & \sigma^2 & \theta \\ \theta & \theta & \theta & \sigma^2 \end{vmatrix}$	$V=\begin{vmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{vmatrix}$

σ_i^2 = variance of responses for time period i; θ_{ij} = covariance of response for periods i and j; when the variances are all equal $\sigma_i^2=\sigma^2$ for all i; in (ii) the covariances depend on the separation of the period, $\theta_{ij}=\theta_{|i-j|}$; with equal covariances $\theta_{ij}=\theta$ for all i and j

unbalanced. As a compromise, data are often analysed using univariate techniques applied either to the mean responses over time or, if there is evidence of a treatment by time interaction, to the individual responses at selected time periods. Neither approach is entirely satisfactory since trials will often lack the power to detect such interactions and calculating a mean value over time takes no account of missing values which are more likely to occur in later time periods. Other summary statistics, such as the slope of the linear regression of response against time, have also been used[24] but do not have general applicability. A more direct solution is to model the variance/covariance structure for responses at successive time periods and estimate the corresponding parameters along with fixed effects. Maximum likelihood provides a general method for this even when the data are unbalanced. Laird and Ware[25] develop the method of maximum likelihood for unbalanced longitudinal studies. Jennrich and Schluchter[26, 27] demonstrate in more detail the application of ML techniques, including REML, in repeated measures studies and discuss the structure of the covariance matrix. We suggest that REML is used in preference to ordinary maximum likelihood to reduce the problem of biased covariance estimates.

Table IV gives possible forms for covariance structure for repeated measures. In the most general case (i), the variances of responses σ_i^2 differ for each time period i and the covariance between periods i and j is θ_{ij}. For the general autoregressive model (ii), the variances are equal and the covariance θ_{ij} between periods depends on their separation $|i-j|$. This may be an appropriate model when time periods are evenly spaced. For the particular case of a first-order autoregressive model, the covariances decrease exponentially with time separation, $\theta_{ij}=\theta^{|i-j|}$. For the uniform covariance model (iii), all covariances are equal. This is equivalent to the model with random patients and uncorrelated errors used for crossover trials where $\theta=\sigma^2\gamma/(1+\gamma)$. Finally, for model (iv), responses are uncorrelated and there are no patient effects.

A drawback of fitting a general covariance structure is the large number of parameters to be estimated. For example, with 10 time periods it would be necessary to estimate 10 variance parameters and 45 covariance parameters, which is not feasible in a small trial with perhaps

Table V. REML estimates of variances and correlations in response for different error models for eczema trial

Error model	Period	Variances	Correlations				
General covariance structure	1	0·95					
	2	0·96	0·74				
	3	0·73	0·52	0·62			
	4	0·76	0·39	0·53	0·67		
	5	0·73	0·53	0·45	0·38	0·45	
	6	1·64	0·53	0·62	0·33	0·44	0·63
Uniform	1–6	0·94	0·48				
Uncorrelated	1–6	0·87	0				
Uniform with separate covariances	1–6 A	1·08	0·56				
	1–6 P	0·78	0·38				

A = active treatment, P = placebo

several missing values. Simpler forms of covariance structure should therefore be considered and tested for suitability.

Finally, the model can be extended to allow separate covariance structures for each treatment group. This would be appropriate, for example, if active treatments are known to be more variable in response than the placebo.

3.1. A parallel-group trial in eczema

Here we consider a parallel group trial to compare the efficacy of a treatment for eczema with a placebo over a six-month period. The severity of eczema is measured by an overall redness score for 20 parts of the body. Measurements were made on each patient before treatment, and then at six time periods after treatment for female patients and at three time periods for male patients. There were 36 women and 14 men in the trial and men missed the first two and the fifth post-treatment assessments. There was a further imbalance because several patients did not complete the study so their scores are only available for the earlier periods. Initially separate analyses were carried out for women and men but no differences were observed so a combined analysis is presented here.

The response model for redness score (RS) at different periods was defined by

$$\log \text{RS} = \text{Initial} \log \text{RS} + \text{Patient effect} + \text{Treatment effect} + \text{Period effect}$$

$$+ \text{Treatment} \times \text{Period effect} + \text{Error},$$

where pre-treatment log redness score was included as a covariate to reduce patient to patient variation. A logarithmic transformation was used for redness score to normalize its distribution. Several different covariance structures for the error term were considered and their parameters estimated using REML. With 50 patients in the trial, tests for comparing alternative covariance structures should have reasonable power.

Initially an unstructured covariance matrix was used. This indicated no systematic pattern in the size of the correlations with separation of time periods (Table V) so an autoregressive model was not considered appropriate. There were, however, differences in variability of response for individual periods; in particular the sixth period was more variable than the others. An alterna-

Table VI. Log-likelihood, degrees of freedom and estimated treatment differences for the eczema trial

Error model	$-2\times$ Log-likelihood	Covariance parameters	Estimated treatment difference	SE
General	566	21	0·402	0·224
Uniform	592	2	0·416	0·218
Uncorrelated	646	1	0·534	0·125
Uniform with separate treatment covariances	590	4	0·402	0·214

tive model assuming a uniform covariance structure was nevertheless found to give no worse a fit than the general model ($\chi^2_{19} = 26$, Table VI). The model with uncorrelated errors gave a significantly worse fit than the uniform covariance model ($\chi^2_1 = 54$). This was unsurprising as the estimated correlation between repeated measures for the uniform model was 0·48, equivalent to $\hat{\gamma} = 0{\cdot}92$, which indicates a substantial component of variance due to patients. Finally a uniform model was fitted with different covariances for each treatment. This gave no improvement over the single covariance model ($\chi^2_2 = 2$) although there was some suggestion that the variance component for the active treatment was higher than that for the placebo.

Table VI gives the estimated treatment difference and its standard error for each model. The estimated standard error for the uncorrelated model is misleadingly small as it is erroneously assumed that responses at different time periods provide independent estimates of treatment effects. Differences among the treatment estimates and standard errors for the other three models is small, suggesting that estimates are fairly robust to precise specification of variance/covariance structure. No significant effect was observed for period or treatment by period interaction.

4. MULTICENTRE TRIALS

So far we have considered the analysis of single centre trials. However, a clinical trial may be carried out at several centres either because insufficient patients are available for the study at any one centre, or with the deliberate intention of assessing the effectiveness of the treatments in several settings. The extra variability in treatment response introduced with a multicentre trial may be taken into account in the analysis by including an effect for centres and treatment by centres interaction in the model. Deciding whether centre effects should be fixed or random in such cases has been the subject of some debate.[28–30] In practice, the choice will depend on the purpose of the experiment and, particularly, whether treatment estimates are to relate only to the set of centres used in the study or, more widely, to the circumstances and locations sampled by the trial centres. In the former case, *local* treatment estimates for individual centres are provided by taking centre and treatment by centre effects as fixed. To predict *global* treatment estimates, centre and treatment by centre effects should be taken as random.

Taking centre effects as random also allows recovery of any between-centre treatment information which will be present when the relative sizes of the treatment groups differ between centres. The amount of extra information will depend on the degree of treatment imbalance across centres (in extreme cases, some treatments may be completely omitted from some centres) and γ, the ratio of between-centres to treatments by centre variance components (see Figure 1).

Including treatment by centre interaction as a random effect in the model will lead to an increase in the standard errors of treatment differences when there is heterogeneity across centres. Conventionally, the decision on whether to include a treatment by centre interaction in term in the model is based on a significance test applied at an arbitrarily specified significance level. When an interaction is present but shows as non-significant, this approach will lead to an underestimate in standard error of treatment differences. We suggest a more appropriate strategy is to include random treatment by centre effects whenever their variance component is positive, as advocated by Patterson and Nabugoomu.[31]

The estimated variance components for the random effects may be of interest in themselves. For example, knowledge of the relative sizes of centre, treatment by centre and patients variance components may be used in trial design to calculate the increase in sample size required to allow for the increased variability expected from sampling several centres. If the variance components are to be interpreted and used with confidence it is important that estimates are based on a reasonable number of observations.

Meta-analyses are increasingly used to combine results from several clinical trials assessing the same treatments so as to provide a more precise overall estimate of the treatment effects. However, the usual Peto estimates,[32] which are based on the assumption of fixed trial effects, give a false impression of accuracy if there is heterogeneity between the trials. This was recognized by DerSimonian and Laird[33] who described the use of a random effects model, including treatment by trials interaction. This provides global treatment estimates and standard errors if the trials are representative of the full range of possible circumstances.[34] When the original data for each trial are available, efficient estimates of the overall treatment effects and variance components for trials and treatment by trials are obtained from a REML analysis in the same way as proposed for multicentre trials. An alternative Bayesian approach which allows for heterogeneity between trials or centres has also been considered.[35–37]

4.1. A multicentre trial in hypertension

Our example comes from a study of three treatments for the management of essential hypertension involving 29 centres. Each patient received one treatment at random and diastolic blood pressure, the main response variable, was recorded before and after treatment. A total of 285 patients completed the study and the distribution of treatment groups across centres is shown in Table VII. At 21 of the centres, treatment groups were unequal, indicating that additional information on treatment comparisons is available for recovery between centres.

The full model for diastolic blood pressure (DBP), including treatment by centre interactions, was taken to be

$$\text{Final DBP} = \text{Initial DBP} + \text{Treatment effect} + \text{Centre effect}$$

$$+ \text{Treatment} \times \text{Centre effect} + \text{Random error},$$

where initial diastolic blood pressure was included as a covariate to reduce between-patient variation.

The model was fitted using REML, initially with treatment by centre interactions omitted. Centre effects were taken first as fixed, then as random (Table VIII). Differences between centres were small and, for the random model, the centres variance component was much smaller than the patients variance component ($\hat{\gamma} = 0{\cdot}035$). The extra treatment information recovered between centres is reflected in the standard errors of treatment means for the random centres model which

Table VII. Number of patients receiving each treatment at each centre in a multicentre trial comparing three treatments for hypertension. Centres with unequal sizes of treatment groups are indicated by*

Centre		Treatment		Total
	A	B	C	
1*	13	14	12	39
2*	3	4	3	10
3*	3	3	2	8
4	4	4	4	12
5*	4	5	2	11
6*	2	1	2	5
7	6	6	6	18
8	2	2	2	6
9*	0	0	1	1
11	4	4	4	12
12*	4	3	4	11
13*	1	1	2	4
14	8	8	8	24
15*	4	4	3	11
18	2	2	2	6
23*	1	0	2	3
24*	0	0	1	1
25*	3	2	2	7
26*	3	4	3	10
27*	0	1	1	2
29*	1	0	2	3
30*	1	0	2	3
31	12	12	12	36
32*	2	1	1	4
35	2	1	1	4
36*	9	6	7	22
37*	3	1	2	6
40*	1	1	0	2
41*	2	1	1	4
Total	100	91	94	285

are on average 10 per cent less than for the fixed centres model and have a smaller range. Since there are over twenty centres, their component of variance can be estimated with reasonable precision and thus a random centres model, which combines between- and within-centres treatment information, is preferred in this case.

We now consider the full model allowing for treatment by centre interaction. Taking these effects as random, the REML estimate of the treatment by centre variance component is of similar size to that for centre (Model 3, Table VIII). Including interaction effects in the model leads to an increase in standard errors of treatment estimates. If a treatment by centre interaction is indeed present in the study, omitting its effects in the model will lead to standard errors of treatment estimates being underestimated.

Finally, if treatment by centre interactions are taken as fixed, we cannot logically estimate overall treatment effects as all treatments do not appear at all centres. This problem is avoided by assuming random treatment × centre effects.

Table VIII. REML estimates for diastolic blood pressure (mmHg) in multicentre hypertension trial

Model	Fixed effects	Random effects
1	Treatment + Centre	—
2	Treatment	Centre
3	Treatment	Centre + Treatment × Centre

	Variance component (standard errors)		
Model	Centre	Treatment × Centre	Patient
1	—	—	73·9 (6·54)
2	2·65 (3·24)	—	75·1 (6·63)
3	1·77 (3·55)	3·51 (4·96)	72·7 (7·00)

	Treatment means (standard errors)		
Model	A	B	C
1	96·1 (1·05)	94·4 (1·10)	92·5 (1·02)
2	95·8 (0·95)	94·4 (0·98)	92·7 (0·97)
3	95·7 (1·01)	94·2 (1·05)	93·0 (1·03)

5. COMPUTER SOFTWARE

REML is now available in SAS,[15] BMDP[16] and Genstat[17] and other more specialized multilevel modelling packages such as ML3[38] and HLM.[39] Although the same estimation procedure is used in all packages the presentation of the output and the terminology varies. More importantly, there is no agreed procedure for significance testing, particularly in small samples. Some of the features of the three packages are summarized in Table IX.

In SAS version 6.07 the mixed model may be fitted using the MIXED procedure which uses REML as the default method. The procedure allows the user to specify the structure of an overall covariance matrix, or separate covariance matrices for particular groups (for example, for different treatments). Estimates of variance components and comparisons among fixed effects are output with their standard errors. Test statistics and *P*-values are provided for the significance of variance components (asymptotic *z*-tests) and for comparing fixed effects (*t*-tests). The overall significance of fixed effects is tested by a Wald *F*-statistic.

In BMDP, REML is provided as an option in the linear model procedures 3V and 5V. BMDP3V is used to fit the general mixed model while 5V is specifically for the analysis of repeated measures data. BMDP5V allows several types of covariance structure to be specified as in SAS. Estimates of variance components and their standard errors are output. Estimates of fixed effects are output but standard errors of differences in effects have to be calculated from a variance/covariance matrix. Wald F or χ^2 test statistics and *P*-values are given for overall tests of fixed effects.

In Genstat, a REML analysis for the mixed model is provided by the VCOMPONENTS and REML directives. Estimates of variance components and fixed effects are output with their standard errors and separate standard errors may be obtained for all differences in fixed effects. Fixed effects may be tested using the asymptotic Wald statistic or REML likelihood ratio test which appears to be more conservative. Although Genstat allows the user to specify a general variance/covariance matrix, the parameters are fixed by the user and cannot be estimated as for the other packages.

Table IX. Features currently available for REML analysis in BMDP, Genstat and SAS

	BMDP (1990 version)	Genstat (Version 5.3)	SAS (Version 6.07)
Procedure/command	3V	VCOMPONENTS and REML	PROC MIXED
Random effect estimates and SEs	Yes	Yes	Yes
Fixed effect differences and SEs	Not directly	Yes	Yes
Significance tests:			
fixed effects	Wald F, t-test	Wald χ^2, LR-test	Wald F, t-test
random effects	comparing deviances	comparing deviances	z-tests
General covariance matrix for repeated measures	Yes (procedure 5V)	No	Yes
Separate variance components for groups	Not directly	Not directly	Yes
Negative estimates of variance components	No	Yes	Yes
Allow missing class levels	Not directly	Yes	Not directly

Significance tests provided by the packages should be interpreted cautiously, particularly when the sample size is small. In particular, we have found that SAS and BMDP give different P-values for testing the significance of fixed effects using the Wald F statistic. This is caused by the different choice of degrees of freedom for the denominator.[13] Genstat gives a likelihood ratio test for fixed effects in REML but the basis for the test has not yet been published. Uncertainty also exists over choice of degrees of freedom for t-tests of individual treatment comparisons. Giesbrecht and Burns[40] suggest how the degrees of freedom might be estimated, but the approach is not used by any of the packages. For random effects only asymptotic tests are provided. We recommend that random effects are tested by comparing deviances (that is, -2 log-likelihoods) of models with and without the effects, as was done in the repeated measures example of Section 3. This is preferable to the z-test provided by SAS which compares each effect with its standard error.

Results should also be interpreted with caution when estimates of variance components are given as negative or very close to zero. Negative estimates are allowed in Genstat and as an option with SAS. A negative estimate can be caused by errors in the data or model specification, but is most likely to indicate that the true variance component is zero. In this case we recommend refitting the model with appropriate random effect term removed. BMDP and SAS (by default) do this automatically when a variance estimate would be negative, and print the estimate as zero together with parameter estimates for the reduced model.

7. DISCUSSION

The main aim of this paper has been to demonstrate how the estimation of treatment effects in clinical trials can be improved by taking full account of the variance/covariance structure of the responses. With unbalanced, multi-classified data, a generalized (weighted) least squares analysis provides efficient treatment estimates by combining information across all classification strata, with weights derived from the stratum variance components. Until now, however, inefficient and somewhat arbitrary methods have been used for estimating these variance components, for example by equating sums of squares.[14] REML is presented here as an efficient maximum likelihood method which can be applied to any general data structure. By contrast to standard maximum likelihood (ML) it gives identical treatment estimates to ANOVA for complete data and for balanced incomplete designs.

Even using efficient REML estimates, variance components used for deriving weighted estimates will still be poorly estimated in small samples. Then, as noted in Section 2, theoretical efficiency gains from recovery of between-stratum information which are based on known weights, may not be achievable. The loss of efficiency from estimating weights was investigated by Yates[2] and, more recently, by Nabugoomu and Allen.[41] Cochran and Cox[42] recommended that interblock information should only be recovered when there is at least 10 degrees of freedom for estimating the block mean square. This was the case for all the examples in this paper and hence the estimated variances of treatment means can be used with some confidence.

Comparison of REML estimates of variance components with ML estimates[43] has shown that REML estimates for block strata can be more variable than ML estimates. However, ML estimates are generally biased downwards relative to the units variance and this bias can be large.[6,44] REML estimates of the standard errors of variance components are generally biased upwards. A useful extension of REML allows different groups of patients (for example, grouped by sex or treatment) to be described by different variance components. This was shown in the repeated measures example where the response of patients in the active treatment group may be more variable than the placebo group. Similarly, in a multicentre trial, the method can be used to allow for treatment differences in the variability of response across centres.

The greatest obstacle at present to the routine use of REML in the analysis of clinical trials is the lack of established procedures for significance testing. In least squares analysis, there is a well developed Normal sample theory underlying the significance tests of fixed effects based on the ratio of 'treatment' and 'error' sums of squares, both of which are χ^2 with known degrees of freedom. The difficulty with inference in REML lies in identifying appropriate degrees of freedom for the error as this is obtained by averaging across a number of strata. SAS and BMDP both provide small sample Wald F-tests and t-tests but they disagree over the appropriate number of error degrees of freedom. When errors are estimated with a large number of degrees of freedom, then asymptotic Wald χ^2 test for overall significance of fixed effects, and z-tests for individual contrasts, can be applied with greater confidence. However, tests based on large sample assumptions are unlikely to be acceptable to drug regulatory authorities unless a general rule can be constructed to indicate when these assumptions are valid.

APPENDIX: MODEL SPECIFICATION AND PARAMETER ESTIMATION

The general mixed linear model can be specified for response variable $y = (y_1 \dots y_n)$ as

$$y = X\alpha + \sum Z_j \beta_j + e \quad (1)$$

where $\alpha' = (\alpha_1 \dots \alpha_t)$ are t vectors of fixed effects with levels $r_1 \dots r_t$, X is an $n \times g$ design matrix with $g = \sum r_i$, and e is a Normally distributed random error variable with mean 0, and variance $I\sigma^2$. The model includes c additional random variables $\beta_1 \dots \beta_c$ where β_j (of length s_j) is Normally distributed with mean 0, variance $I\sigma_j^2$, and has $n \times s_j$ design matrix Z_j.

The mean and variance/covariance matrix for y can be expressed as

$$E(y) = X\alpha,$$

and

$$\text{var}\,(y) = V = \sigma^2\,[I + \sum Z_j Z_j' \gamma_j], \tag{2}$$

where $\gamma_j = \sigma_j^2/\sigma^2$. (For the repeated measures model, the errors e in (1) are not independent and this leads to a different specification for the variance matrix V.)

We wish to estimate the fixed effects $\alpha_1 \dots \alpha_t$, the unit variance σ^2 and variance ratios $\gamma_1 \dots \gamma_c$ for random effects. If we know the variance matrix V we can use generalized least squares and estimate the fixed effects as

$$\hat{\alpha} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

with

$$\text{var}\,(\hat{\alpha}) = (X'V^{-1}X)^{-1}. \tag{3}$$

To estimate the variance components we could use maximum likelihood but this takes no account of the information used in estimating fixed effects and so leads to biased treatment estimates. REML takes account of this loss of information by modifying the likelihood equation to exclude the contribution from fixed effects. The likelihood then becomes

$$L = -1/2\,[\log|V| + (y - X\hat{\alpha})'V^{-1}(y - X\hat{\alpha}) - \log|X'V^{-1}X|^{-1}], \tag{4}$$

where $\hat{\alpha}$ are the estimated fixed effects from (3). Note that the REML equation differs from the classical likelihood equation by a term $\log|X'V^{-1}X|^{-1}$ associated with the variance of treatment estimates. REML estimates of the variance parameters of V may then be obtained by maximizing (4) using an iterative scheme.[5]

The full REML method thus consists of estimating fixed effects by generalized least squares (3) and random effects by residual maximum likelihood (4). The difference between REML and most other procedures lies in the method of estimating variance components. A good review of alternative methods is given by Searle *et al.*[7] The various ANOVA variance estimators used by other authors[2,3,14,21] are equivalent to REML estimators for complete data and balanced incomplete block designs. Another family of variance estimators proposed by Rao[45,46] are the minimum norm quadratic unbiased estimators (MINQUE). Under normality assumptions iterating the MINQUE equations once again gives the REML estimators.[6,47] Finally, when there is only one variance component (the unit variance) REML reduces to simple least squares analysis.

ACKNOWLEDGEMENTS

We are grateful to Olivier David for deriving the formulae in Table I and to Desmond Patterson and Fabian Nabugoomu for many helpful discussions.

REFERENCES

1. Fisher, R. A. *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.
2. Yates, F. 'The recovery of inter-block information in balanced incomplete block designs', *Annals of Eugenics*, **10**, 317–325 (1940).
3. Henderson, C. R. 'Estimation of variance and covariance components', *Biometrics*, **9**, 226–252 (1953).
4. Hartley, H. O. and Rao, J. N. K. 'Maximum likelihood estimation for the mixed analysis of variance model', *Biometrika*, **54**, 93–108 (1967).
5. Patterson, H. D. and Thompson, R. 'Recovery of inter-block information when block sizes are unequal', *Biometrika*, **58**, 545–554 (1971).
6. Patterson, H. D. and Thompson, R. 'Maximum likelihood estimation of components of variance', *Proceedings of the* 8th *International Biometric Conference*, 197–207 (1975).
7. Searle, S. R., Casella, G. and McCulloch, C. E. *Variance Components*, Wiley, New York, 1992.
8. Thompson, R. 'The estimation of heritability with unbalanced data. I. Observations available on parents and offspring', *Biometrics*, **33**, 485–495 (1977).
9. Meyer, K. 'Restricted maximum likelihood to estimate genetic parameters in practice' in Dickerson, G. E. and Johnson, R. K. (eds) *Proceedings of the* 3rd *World Congress on Genetics applied to Animal Production Vol XII*, University of Nebraska, 1986, pp. 454–459.
10. Talbot, M. 'Yield variability of crop varieties in the UK', *Journal of Agricultural Science, Cambridge*, **102**, 315–321 (1984).
11. Robinson, D. L. 'Estimation and use of variance components', *The Statistician*, **36**, 3–14 (1987).
12. Chakravorti, S. R. and Grizzle, J. E. 'Analysis of data from multiclinic experiments', *Biometrics*, **31**, 325–338 (1975).
13. Berk, K. 'Computing for incomplete repeated measures', *Biometrics*, **43**, 385–398 (1987).
14. Chi, E. M. 'Recovery of inter-block information in crossover trials', *Statistics in Medicine*, **10**, 1115–1121 (1991).
15. SAS Institute Inc., SAS Campus Drive, Cary, NC 27513, U.S.A.
16. BMDP Statistical Software Inc., 1440 Sepulveda Boulevard, Suite 316, Los Angeles, CA 90025.
17. Genstat Numerical Algorithms Group Ltd., Mayfield House, 256 Banbury Road, Oxford, UK.
18. Gough, K. 'Letter to the editor', *Statistics in Medicine*, **8**, 891–892 (1989).
19. Hafner, J. B., Koch, G. G. and Canada, A. T. 'Some analysis strategies for three-period crossover designs with two treatments', *Statistics in Medicine*, **7**, 471–481 (1988).
20. Jones, B. and Kenward, M. G. *Design and Analysis of Crossover Trials*, Chapman and Hall, London, 1989.
21. Laird, N. M., Skinner, J. and Kenward, M. G. 'An analysis of two-period cross-over designs with carryover effects', *Statistics in Medicine*, **11**, 1967–1979 (1992).
22. Mead, R. *The Design of Experiments*, Cambridge University Press, 1988.
23. Abeyasekera, S. and Curnow, R. N. 'The desirability of adjusting for residual effects in a crossover design', *Biometrics*, **40**, 1071–1078 (1984).
24. Rowell, J. G. and Walters, R. E. 'Analysing data with repeated observations on each unit', *Journal of Agricultural Science, Cambridge*, **87**, 423–432 (1976).
25. Laird, N. M. and Ware, J. H. 'Random effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1982).
26. Jennrich, R. I. and Schluchter, M. D. 'Unbalanced repeated-measures models with structured covariance matrices', *Biometrics*, **42**, 805–820 (1986).
27. Schluchter, M. D. 'Analysis of incomplete multivariate data using linear models with structured covariance matrices', *Statistics in Medicine*, **7**, 317–324 (1988).
28. Fleiss, J. L. 'Analysis of data from multiclinic trials', *Controlled Clinical Trials*, **7**, 267–275 (1986).
29. Grizzle, J. E. 'Letter to the editor', *Controlled Clinical Trials*, **8**, 392–393 (1987).
30. Senn, S. J. and Hildebrand, H. 'Crossover trials, degrees of freedom and its dual', *Statistics in Medicine*, **10**, 1361–1374 (1991).
31. Patterson, H. D. and Nabugoomu, F. 'REML and the analysis of series of crop variety trials', *Proceedings of the XVIth International Biometric Conference*, 77–93 (1992).
32. Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, P. 'Beta blockade during and after myocardial infarction: an overview of the randomised trials', *Progress in Cardiovascular Diseases*, **XXVII**, 335–371 (1985).

33. DerSimonian, R. and Laird, N. 'Meta-analysis in clinical trials', *Controlled Clinical Trials*, **7**, 177–188 (1986).
34. Thompson, S. G. and Pocock, S. J. 'Can meta-analyses be trusted?', *Lancet*, **338**, 1127–1130 (1991).
35. Skene, A. M. and Wakefield, J. C. 'Hierarchical models for multicentre binary response studies', *Statistics in Medicine*, **9**, 919–929 (1990).
36. Louis, A. L. 'Using empirical Bayes methods in biopharmaceutical research', *Statistics in Medicine*, **10**, 811–829 (1991).
37. Carlin, J. B. 'Meta-analysis for 2 × 2 tables: A Bayesian approach', *Statistics in Medicine*, **11**, 141–158 (1992).
38. ML3, Multi-level Models Project, 11 Woburn Square, London, WC1H 0NS.
39. HLM, Scientific Software Incorporated, 15-25 East 53rd Street, Suite 906, Chicago, Illinois 60615, U.S.A.
40. Giesbrecht, F. G. and Burns, J. C. 'Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results', *Biometrics*, **41**, 477–486 (1985).
41. Nabugoomu, F. and Allen, O. B. 'The estimation of fixed effects in a mixed linear model', *Proceedings of the* 1993 *Kansas State University Conference on Applied Statistics in Agriculture*, (1994).
42. Cochran, W. G. and Cox, G. M. *Experimental Designs*, 2nd edition, Wiley, New York, 1957.
43. Corbeil, R. R. and Searle, S. R. 'A comparison of variance component estimators', *Biometrics*, **32**, 779–791 (1976).
44. Harville, D. A. 'Maximum likelihood approaches to variance components estimation and to related problems', *Journal of the American Statistical Association*, **72**, 320–338 (1977).
45. Rao, C. R. 'Estimation of variance and covariance components – MINQUE theory', *Journal of Multivariate Analysis*, **1**, 257–275 (1971).
46. Rao, C. R. 'Estimation of variance and covariance components in linear models', *Journal of the American Statistical Association*, **67**, 112–115 (1972).
47. Hocking, R. R. and Kutner, M. H. 'Some analytical and numerical comparisons of estimators for the mixed A.O.V. model', *Biometrics*, **31**, 19–27 (1975).

The Edinburgh randomised trial of breast cancer screening: results after 10 years of follow-up

F.E. Alexander[1], T.J. Anderson[2], H.K. Brown[1], A.P.M. Forrest[3], W. Hepburn[1], A.E. Kirkpatrick[4], C. McDonald[3], B.B. Muir[4], R.J. Prescott[1], S.M. Shepherd[1], A. Smith[1] & J. Warner[5]

[1]Department of Public Health Sciences, [2]Department of Pathology and [3]Scottish Cancer Trials Office, The University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK; [4]Edinburgh Breast Screening Clinic, Springwell House, 26 Ardmillan Terrace, Edinburgh EH11 2JL, UK; [5]Information and Statistics Division, Scottish Health Services Common Services Agency, Trinity Park House, South Trinity Road, Edinburgh EH5 3SQ, UK.

Summary The Edinburgh Randomised Trial of Breast Cancer Screening recruited 44,288 women aged 45–64 years into the initial cohort of the trial during 1978–81, and 10 years of follow-up is now complete. A total of 22,944 women were randomised into the study group and were offered screening for 7 years; the remaining women formed the control group. After 10 years, breast cancer mortality is 14–21% lower in the study group than in the controls depending on the precise definition of the end point. These differences are not statistically significant; for breast cancer as the underlying cause of death the relative risk is 0.82 (95% confidence interval 0.61–1.11). Rates of locally advanced and metastatic cancer were substantially lower in the study group, but screening has failed to achieve marked reductions in rates of small node-positive cancers. Those women who accepted the final invitation to screening have been monitored over the 3 year period prior to their first screen under the UK service screening programme. Interval cases, expressed as a proportion of the control incidence, increased from 12% in the first year to 67% in the third year. The reduction in breast cancer mortality for older women (aged at least 50 years) is the same as that for the total study group for this duration of follow-up. For analyses of breast cancer mortality in younger women updates recruited to the trial from 1982 to 1985 (10,383 women with 6–8 years' follow-up) have been included. The reduction in breast cancer mortality for women aged 45–49 years at entry was 22% (relative risk = 0.78, 95% confidence interval = 0.46–1.31).

The Edinburgh Randomised Trial of Breast Cancer Screening (Roberts *et al.*, 1984) was started in 1978. A total of 44,288 women in Edinburgh were randomised into two groups of approximately equal size. A total of 22,944 women entered the study arm of the trial from 1978 to 1981; these women were invited to participate in a screening programme that included seven annual screens by clinical examination (for seven consecutive years) and mammography (at the first screen and at 2 yearly intervals). These same women formed the Edinburgh component of the study population of the Trial of Early Detection of Breast Cancer Screening or TEDBC (UK Breast Cancer Detection Working Group, 1981). Mortality results for the TEDBC for 10 years of follow-up have recently become available (UK Breast Cancer Detection Working Group, 1993) and their data include comparisons of the present study group with geographical controls. The control groups for the TEDBC and the Edinburgh trial are entirely distinct.

The first report of the Edinburgh trial (Roberts *et al.*, 1990) included results for 7 years' follow-up for breast cancer mortality and 5 years' follow-up for breast cancer incidence. A reduction of 17% in breast cancer mortality in the study group was observed at that time (relative risk = 0.83, 95% confidence interval 0.58–1.18). The results of 10 years of follow-up are now reported. In addition to the initial cohort of the trial, younger women (aged 45–49 years) were entered annually from 1982 to 1985; the results of shorter periods of follow-up for these women have been included here.

Following the publication of the Forrest (1987) report, mammographic screening for breast cancer was introduced into National Health Service policy in 1988. Invitations to service screening for women who had been screened during the trial period were scheduled so that, so far as possible, they were screened during their tenth year of follow-up and 3 years after their last trial screen. In this way we are able to provide the first estimates from a UK population of the frequency of interval cases for women in regular screening whose screens are scheduled to occur at 3 yearly intervals.

This report focuses on three topics: breast cancer mortality after 10 years of follow-up, the effect of screening women aged under 50 years and the consequences of a 3 year inter-screening interval.

Methods

The trial population

Detailed methods have been described previously (Roberts *et al.*, 1984). The geographical base for the trial comprised 87 general practices within the city of Edinburgh. These were enrolled in turn between June 1979 and December 1981 (September 1978 for one practice). As each practice joined the trial all women aged between 45 and 64 years and on the practice list were admitted to form the initial cohort of the trial. Women who attained the age of 45 after the practice entry date and others (over 45 years of age) who moved into the study area were entered from 1982 to 1985. In order to maximise the numbers available for subgroup analyses of younger women (aged 45–49 years at entry) these later entrants have been included in these analyses and are referred to as updates.

In 1985, women in the trial were flagged with the General Registry Office in Edinburgh to obtain information on cancer incidence and death. This ensures follow-up even for women who have moved away from Edinburgh. The present report is restricted to women who were successfully flagged (97% of the total).

Women who had breast cancer diagnosed before entry to the trial are ineligible, but this was often established retrospectively following receipt of a death notification (see below).

Correspondence: F.E. Alexander.
Received 8 November 1993; and in revised form 18 February 1994.

Randomisation and screening

The 87 practices were randomised to study or control status, which provided cluster randomisation for individual women who derived their status within the trial from that of their practice at entry. Women in the study group were offered screening, and those who attended (61.3%) underwent two-view mammography and clinical examination at their initial visit (prevalence screen). Further screening (incidence screens) used annual clinical examination for 6 years and included single-oblique view mammography in alternate years. Attendance rates fell with time and were just over 50% during the final (seventh) year of fieldwork (Roberts *et al.*, 1990).

For the majority of women who continued in screening each screen occurred within 1 year of the intended time – so that, for example, the screen in 'year 3' occurred between 2 and 3 years from survey entry; these women are described as having 'regular' screening. The NHS introduced service screening in Scotland in 1988 for women then aged 50–64 years. For practical and administrative reasons this had to be introduced gradually and the Edinburgh programme was coordinated with the present trial in such a way that all women who were still in regular screening had their first invitation to service screening at (approximately) 3 years after their last (year 7) trial screen; this was during their tenth year of follow-up.

Follow-up

Apart from the collection of medical information and screening histories at the screening clinic, follow-up of women in the two arms of the trial has been identical (Roberts *et al.*, 1984, 1990). For the field-work period of the trial (1978–88) local follow-up for both breast cancer incidence and total mortality were used as independent data sources alongside flagging. Since then, flagging has provided the primary, and for mortality data the only, source of follow-up. Scrutiny of counts of death notifications from flagging suggests that these data are virtually complete after a time lag of 6 months. Data for the present analysis were finalised in January 1993 so that ascertainment of relevant deaths (occurring 1991 or earlier) could be ensured.

Whenever a death certificate mentioned breast cancer as a cause of death for a woman who had not already been identified as a breast cancer case the trial staff sought confirmation of diagnosis. If this occurred before survey entry date the woman was ineligible for the trial.

Flagging for cancer incidence may be less reliable than flagging for mortality. Therefore the entire trial cohort was matched against the Scottish Cancer Registration Scheme database held centrally by the Information and Statistics Division (ISD) of the Scottish Health Services Common Services Agency. The cancer registration database is matched annually with the Scottish hospitals inpatient database. Linkage with the trial database used probability matching and included all notifications of cancer registered up to the end of 1991. These methods optimised ascertainment of breast cancer incidence in the trial population for the full 10 years of follow-up.

Analysis of breast cancer mortality

The primary end point for analysis is 'breast cancer mortality' and, in the initial report (Roberts *et al.*, 1990), this was defined to be mention of breast cancer on either part 1 or part 2 of the death certificate. We have continued to adopt this definition here but have introduced two alternatives. Coding of death certificates in Scotland permits one cause to be *underlined* as the underlying cause of death, and this procedure follows WHO rule 3 informally but does not use the systematic approach adopted in England and Wales (OPCS, 1985). Whenever breast cancer was underlined on the death certificate the death has been classified as a breast cancer death in the present analyses. We have checked all remaining deaths of women with breast cancer diagnosed during the trial period to derive two classifications of breast cancer as *underlying* cause of death:

Definition 1: breast cancer was the underlined cause of death on death certificate or formal application of WHO rules (OPCS, 1985) attributed the death to breast cancer when a non-specific cause was underlined.

Definition 2: breast cancer was the underlined cause of death on death certificate or another cause was underlined but case note review identified breast cancer as the underlying cause.

For definition 2 doubtful cases were considered by a committee of three doctors. This was more accurate but subject to potential bias and definition 1 has been taken as the principal definition of the end point.

As in the previous report, deaths occurring as results of non-epithelial cancers in the breast (e.g. sarcoma, lymphoma) are not included in the analysis.

Following the trial protocol (Roberts *et al.*, 1984), the main analysis has been of breast cancer mortality in the whole of the initial cohort for 10 years from survey entry. We also report the results of two subsidiary analyses focused on possible differential effects of screening women over and under 50 years of age. Firstly, breast cancer mortality over the 10 year period for the initial cohort has been analysed separately for the two age groups: 45–49 years and 50–64 years at survey entry. Secondly, we have included updates in the analysis of younger women; women entered during 1982–83 are followed-up for 8 years and later updates for 6 years (since randomisation was in two groups by year at entry). The analyses to be conducted were decided in advance of data inspection.

For the statistical analysis mortality has been expressed as rates of breast cancer deaths per 10,000 woman–years at risk. Rates in study and control practices were compared and their ratio calculated. As before (Roberts *et al.*, 1990), a modified logistic regression procedure incorporated adjustment for extrabinomial regression (Williams, 1982) so as to respect the cluster randomisation. All analyses were implemented in 'GLIM' and stratified by age at survey entry (45–49, 50–54, 55–59 and 60–64 years). Where updates are included there has been further stratification by length of follow-up. Cumulative breast cancer mortality curves are expressed as rates per 10,000 women entering the study but are adjusted to take account of women–years at risk. The general practice 'clusters' have been classified into three groups by levels of a socioeconomic score (SEG) as in the previous report (Roberts *et al.*, 1990).

Analysis of breast cancer incidence

Staging of disease follows standard UICC clinical staging (UICC, 4th edn, 1987) and stage 0 corresponds to carcinoma *in situ*; for cases which are stage I–II pathological information is also provided.

Cancers detected in women who had attended their last trial screen (i.e. seventh screen, 6–7 years from entry) but had not yet had a service screen are described as 'intervals'; cases arising in the period 36–42 months after the last trial screen and before invitation to the service screening are included as intervals. The *proportional incidence of interval cancers* is the ratio of the incidence rate of these interval cancers to that observed in the control group (adjusting for the age distribution of the population at risk).

Cumulative incidence of cancers known to be advanced in the total trial population have been plotted by year of follow-up for two definitions of 'advanced': firstly, UICC stages III and IV and, secondly, that used by Tabar *et al.* (1985). These figures are adjusted for women–years at risk.

Results

The trial population is shown in Table I with women-years of follow-up. Of the study population 61.3% responded to the initial invitation to screening, but only 44.1% attended the seventh screen. The numbers of women known to be ineligible on account of pretrial breast cancer are also shown; prospective ascertainment is more complete for the study group and, to avoid consequential bias, these women have been retained in the calculations of women–years for the denominators.

Breast cancer mortality

There have been 250 deaths in women with breast cancer diagnosed during the trial period, including 196 (78.4%) cases in which breast cancer was the underlined cause on the death certificate and which were classified as breast cancer deaths in all analyses. Case notes were reviewed for a random sample of 50 of these women, with the cause confirmed in 47; three deaths were not attributable to breast cancer.

The three definitions of breast cancer death agreed, positively (199) or negatively (23), for 222 (89%) of the 250 deaths, and the two definitions of underlying cause agreed for 232 (94%). Patients in whom doubt about cause of death was noted included just two of the younger women (aged 45–49 years).

Details of breast cancer mortality for the total initial cohort are provided in Table II. The mortality rates are lower for the study population for all definitions with reductions of 14–21%. None of the results achieved statistical significance, and the confidence intervals are wide. For the principal end point (breast cancer as the underlying cause using death certificate information) the reduction was 18% (relative risk = 0.82, 95% confidence interval 0.61–1.11). Cumulative breast cancer mortality by year of follow-up is shown in Figure 1a.

Results for subgroup analyses using the principal end point definition are reported in Table III. These provide no evidence that a larger mortality reduction has been achieved in older women (i.e. those aged 50 years or more at trial entry). When the younger women were analysed separately (Table III and Figure 1b) the total numbers of breast cancer deaths were small and estimates of benefit consequently imprecise. The estimated mortality reductions are similar to those for the entire study group. When the calculations

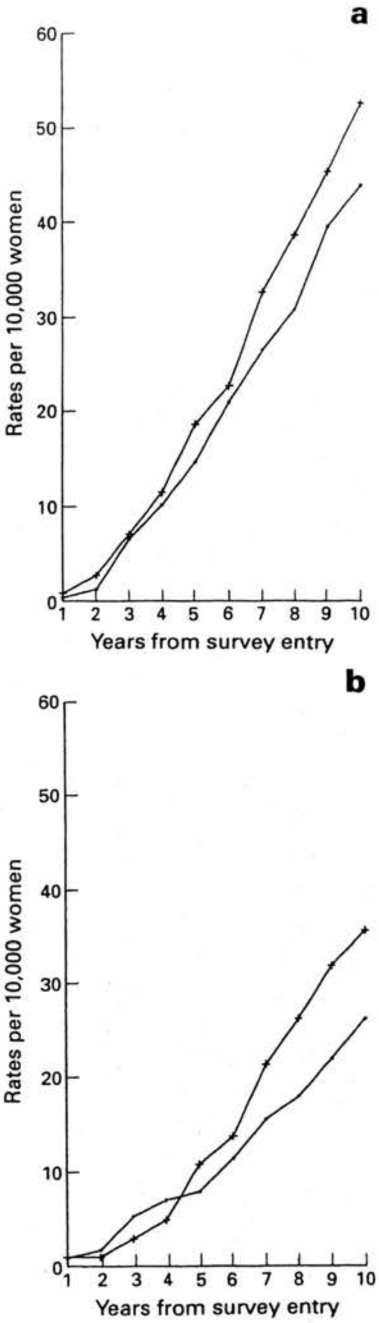


Figure 1 Cumulative mortality from breast cancer in study (●) and control (+) groups over 10 years of follow-up: (**a**) all ages; (**b**) aged 45–49 years at survey entry and including updates. Breast cancer death refers to the underlying cause from death certificate information (definition 1: see Methods section).

Table I Trial population[a] and women–years of follow-up

	Study population		*Control population*	
Age at entry	*Number of women*[b]	*Women–years of follow-up*	*Number of women*[b]	*Women–years of follow-up*
45–49	5795 (59)	56750	5596 (32)	54588
50–54	5878 (65)	57021	5168 (40)	49603
55–59	6109 (112)	57993	5749 (56)	53872
60–64	5162 (102)	47451	4831 (49)	43758
Total	22944 (338)	219215	21344 (177)	201821

[a]These figures are for the initial cohort defined in the Methods section and exclude women who were not successfully flagged; corresponding figures for the updates aged 45–49 at entry are: study population, 5,710 women (40,456 women–years); control population, 4,673 women (34,178 women–years). [b]Figures in brackets are counted of women known to be ineligible because of prior breast cancer diagnosis. Eligibility of screened women was established prospectively, but ineligibility of control women and non-attenders is often established retrospectively after death (from breast cancer) has occurred.

Table II Breast cancer mortality in the initial cohort during 10 years of follow-up

Definition of breast cancer death	*Trial group*	*Number of breast cancer deaths*	*Mortality rate/10,000 women–years at risk*	*Odds ratio (95% confidence interval)*
Death certificate[a]	Study	105	4.79	0.79
	Control	120	5.95	(0.60–1.05)
Underlying cause				
Definition 1[b]	Study	96	4.38	0.82
	Control	106	5.25	(0.61–1.11)
Definition 2[c]	Study	101	4.61	0.86
	Control	108	5.35	(0.66–1.13)

[a]Mentioned as a cause of death on the death certificate. [b]Breast cancer the underlying cause of death from death certificate data (either the underlined cause or derived from formal application of WHO rule 3). [c]Breast cancer underlined cause of death on the death certificate or confirmed as underlying cause – case note review.

reported in Table III were repeated for the other definitions of breast cancer death, the results were qualitatively similar. The mortality reduction for these women did not depend on whether the age at diagnosis exceeded 50 years.

Breast cancer incidence

Altogether, 489 breast cancers were diagnosed in the 10 year period in the study population (22.4/10,000 women–years) and 400 in the controls (20.0/10,000 women–years).

The UICC stage distribution is more favourable for the study population than for the controls (Table IV); percentages of invasive cancers classified as stages III or IV are 17% and 32% respectively. Cumulative rates (Figure 2a) in the control group have always exceeded those in the study group with a 2-fold excess after 10 years. Pathological classifications of stage I and II cancers by size and node status confirm the generally favourable characteristics of cancers in the study group. The cancers are smaller in the study group and, for each size, the proportions of node-positive cancers are lower.

Tabar *et al.* (1985) have combined clinical and pathological data to define a poor-prognosis tumour category (stage III or IV, pathological size $>$20 mm or node positive). The proportions of invasive cancers which are in this category are 53.5% and 73.5% for study and control groups respectively; for the never-screened women the percentage is 80%. However, approximately half (152 or 47%) of all invasive cancers in ever-screened women are poor prognosis according to this definition, and of these 41 (27%) are small node-positive cancers. Cumulative rates of poor-prognosis cancers (Figure 2b) show excesses in the study group for the first 7 years, but thereafter a divergence in favour of the study group is emerging.

The *rates* of interval cases in the 3 years after the trial screening ended (Table V) are compared with rates in the control population the proportional incidence of interval cases increases markedly with time since the last screen from 12% in the first year to 67% in the third year.

Treatment

Altogether, 55% of women with operable invasive cancer were treated with adjuvant systemic therapy. This percentage increased with time from 36% in those diagnosed in study years 1–5 to 79% for years 6–10. These percentages did not differ between the two arms of the trial: 53% of the study group and 57.5% of the control group received adjuvant therapy. Corresponding figures for the updates were 63.5% for the study group and 70.3% for the control group.

All-cause mortality and effect of socioeconomic status

All-cause mortality in the total trial population was 103.9/10,000 women–years, which is similar to that expected in a cohort of Scottish women of this age. The rates in the study group were 15% lower than in the controls (relative risk = 0.85, 95% confidence interval 0.79–0.92). This difference cannot be attributed to breast cancer, which represented only 4.7% of all deaths. When the trial population was split into three groups by SEG of the general practice clusters, the rates were 84.9, 104.1 and 126.1 with the lowest rates in the highest SEG; the trend is statistically significant ($P<$ 0.00001). More women in the study population were in the

Table III Breast cancer mortality[a]: further analyses

Population studies	*Age at entry (years)*	*Follow-up* *First year*	*Last year*	*Trial group*	*Number of breast cancer deaths*	*Mortality rate/10,000 women–years*	*Odds ratio (95% CI)*
Initial cohort	50–64	1	10	Study	79	4.86	0.85
				Control	85	5.77	(0.62–1.15)
Initial cohort	45–49	1	10	Study	17	3.00	0.77
				Control	21	3.85	(0.37–1.62)
Initial cohort				Study	25	2.57	0.78
and updates	45–49	1	6,8,10[b]	Control	31	3.49	(0.46–1.31)

[a]Underlying cause of death derived from death certificate data. [b]Follow-up period available depending on entry year (see Methods section).

Table IV Classification of breast carcinoma in study and control populations

UICC stage	*Pathological classification*	*Study population* *Ever screened*[a]	*Never screened*	*Total*	*Control population Total*
0	TIS	39	3	42	11
I/II	⩽10 mm				
	N_0	57	3	60	27
	N_1	9	2	11	6
	Total[b]	73	7	80	41
	11–20 mm				
	N_0	84	14	98	43
	N_1	32	7	39	28
	Total[b]	126	24	150	85
	21–50 mm				
	N_0	54	11	65	45
	N_1	33	13	46	37
	Total[b]	88	27	115	96
	Size unknown	5	5	10	7
III		16	27	43	76
IV		7	24	31	49
Total[c]		364	125	489	400

[a]These women accepted at least one invitation to screening during the trial period. [b]Totals include those for whom node status is unknown. [c]Totals include those for whom UICC stage is unknown.

highest SEG group (percentages of women–years were 53% and 26% in the study and control groups respectively) and fewer were in the lowest SEG group (27% compared with 42%). Thus the differences in all-cause mortality can be explained at least in part by socioeconomic classifications. It was not possible to calculate an overall RR for breast cancer mortality with adjustment for practice SEG because of a statistically significant interaction ($P<0.005$) between arm of the trial and SEG. Since SEG is applied to GP practices rather than individual women, the imbalance between the two arms of the trial is independent of age.

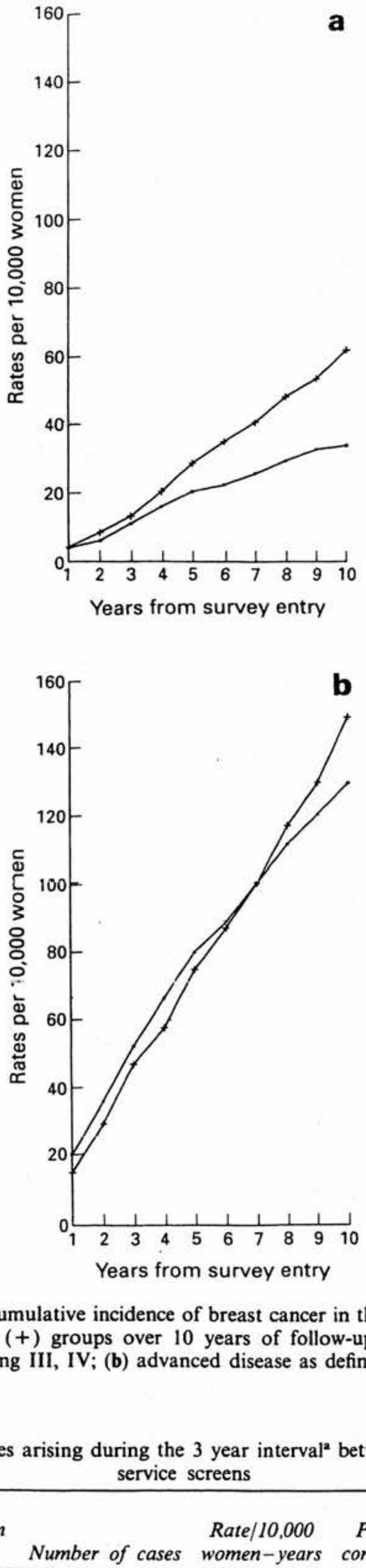


Figure 2 Cumulative incidence of breast cancer in the study (●) and control (+) groups over 10 years of follow-up: (**a**) UICC clinical staging III, IV; (**b**) advanced disease as defined by Tabar *et al.*, 1985.

Table V Cases arising during the 3 year interval[a] between trial and service screens

Times from negative screen (months)	*Number of cases*	*Rate/10,000 women–years*	*Proportion of control incidence*
0–11	2	2.0	0.12
12–23	7	7.1	0.42
24–35	11	11.3	0.67
36 or over	4	11.1	0.65
Total	24	7.3	0.43

[a]This is restricted to women who were still in regular screening in year 7 of the trial.

Discussion

These results based on 10 years of follow-up confirm our earlier findings (Roberts *et al.*, 1990). There is a reduction in breast cancer mortality of around 18% in the total study population. This is not statistically significant and the confidence intervals are wide so that, by themselves, the results are inconclusive. They should be interpreted in context and are consistent with the consensus (Wald *et al.*, 1991) that mammographic screening reduces breast cancer mortality but by rather less than the 30% originally found in the HIP study (Shapiro *et al.*, 1982) and later by the Swedish two-counties trial (Tabar *et al.*, 1985). Other trials in Sweden have achieved smaller reductions, and the recent Swedish overview (Nystrom *et al.*, 1993) which includes all the Swedish trials estimates the reduction at 12 years of follow-up to be 22%; this is based on 882 deaths and is highly statistically significant. No other randomised trials of mammographic screening using an unscreened control group have been conducted outside Sweden since the HIP study, which was started in 1967. The Edinburgh trial, though numerically small, provides useful indications that the benefits observed in Sweden can be achieved elsewhere. This is important since the breast cancer experience in Sweden is very different; for example, cumulative mortality in the Edinburgh *study group* after 10 years (4.38/10,000 women-years) exceeded that in the Swedish two-counties *control group* after 12 years (4.15/10,000 women–years, from Nystrom *et al.*, 1993).

The TEDBC is larger than the Edinburgh trial but is a geographical comparison. Our data are not independent since the Edinburgh study group contributes to the TEDBC; the controls are, however, entirely distinct. Estimated reductions in breast cancer mortality for the TEDBC of 20% have recently been reported (UK Breast Cancer Detection Working Group, 1993); these are statistically significant but rely on difficult adjustments using pretrial breast cancer standardised mortality ratios (SMRs) for the geographical regions.

A number of design or execution problems for the study have been discussed previously (Roberts *et al.*, 1990) but should be noted here; these include the attendance rate, mammographic quality in the early years, loss of efficiency from cluster rather than individual randomisation and statistical power. All of these will have a conservative effect. Other potential sources of bias were, firstly, different use of adjuvant systemic therapy (Early Breast Cancer Trialists Collaborative Group, 1992) between the two groups and, secondly, differential errors of ascertainment of breast cancer deaths. We have demonstrated that the first does not apply and our methods of ascertainment avoid the latter.

Establishing whether death is attributable to breast cancer is difficult in a small number of cases. The review of a random sample of deaths with breast cancer underlined on the death certificate has confirmed previous findings: overestimation of breast cancer as a cause of death of around 6% (Brinkley *et al.*, 1984; UK Breast Cancer Detection Working Group, 1991). The review of case notes reveals that small errors have occurred in the classification of the remaining deaths of breast cancer patients, but these data are compatible with other recommendations (Brinkley *et al.*, 1984) that analyses should use death certificate classifications of the underlying cause of death. Non-differential misclassification in the data will be a further source of conservative bias.

An unexpected consequence of the use of cluster sampling is the bias between the two arms of the trial, which is evidenced in the differences in all-cause mortality. This is in part explained by lower socioeconomic status in the control group (Alexander *et al.*, 1989). There is a body of evidence demonstrating that the direction of the association of socioeconomic status with breast cancer is opposite to that for all-cause mortality (Tomatis *et al.*, 1990; Scottish Breast Screening Programme, 1993). This suggests that the bias should be conservative, but our data have not permitted formal adjustment of the analyses for the effect of socioeconomic status. Despite the possibility of confounding in this study and inadequacy of SMR adjustment in the TEDBC, the point estimates from the two studies are very similar.

We conclude that our findings support the prevailing view that mammographic screening does reduce mortality from breast cancer – at least in women screened when 50 years and over (Wald *et al.*, 1991). Our data do not indicate that this age range fared better than the younger women.

There is currently no consensus on the merits of screening younger women (Beral, 1993; Elwood *et al.*, 1993; Fletcher *et al.*, 1993). The principal difficulty is that no trial of traditional design has been conducted with sufficient statistical power to analyse this age group separately. In addition, most authors fail to distinguish between the short and longer follow-up periods. For short follow-up periods (i.e. 7 years or less) all published trial results give point estimates of the relative risk which are either close to unity or exceed it (Shapiro *et al.*, 1982; Miller *et al.*, 1992; Nystrom *et al.*, 1993). This is supported by the Malmo trial (Andersson *et al.*, 1988) and the only relevant case-control study (Verbeek *et al.*, 1985). For longer follow-up periods the HIP trial reported reductions in breast cancer mortality for younger women (first screened when 40-49 years) which was eventually of the same magnitude as that in older women (Shapiro *et al.*, 1988) and, in one analysis (Chu *et al.*, 1988), achieved statistical significance. This is the only trial with follow-up exceeding 12 years but the Swedish trials have now reported a reduction of 13% after 12 years. Similar results have been reported by the Breast Cancer Demonstration Project, although this latter is neither population based nor randomised (Morrison *et al.*, 1988).

The Edinburgh results are consistent with the emerging consensus that modest reductions in breast cancer mortality are achieved for women first screened when under 50 years, but the benefits appear later than those for older women. At 7 years breast cancer mortality for women in the youngest age group was almost identical for the two arms of the trial. Now there is a reduction of 22% in the study group, but this does not approach statistical significance. These results with a randomised design and independent controls confirm the findings of the TEDBC [relative risk = 0.74 (0.54-1.0)] (UK Breast Cancer Detection Working Group, 1993). The data cannot be interpreted as evidence for population screening for younger women; they do, however, indicate that the optimal age to commence screening remains unknown; this is an important public health question which requires resolution by further randomised trials.

The classification of breast cancers by clinical stage and by pathological size and node status reveals an encouraging difference between study and control groups but is disappointing in one critical respect. This is the failure of screening to reduce the incidence of poor-prognosis disease as defined by Tabar *et al.* (1985) and, in particular, to achieve major reductions in the frequency of nodal metastases amongst the smaller cancers ($\leqslant 20$ mm). Half of all invasive cancers in ever-screened women are poor prognosis compared with 33% in the Swedish two-counties trial (Tabar *et al.*, 1985). In both trials the percentages in the never-screened women (80% and 73% respectively) exceeded those in the controls, but the impact on the total study group is greater in Edinburgh on account of the lower attendance rate. The cumulative incidence of this poor-prognosis disease is not substantially reduced in the Edinburgh study group – although the results for the years 8-10 give preliminary indications that the curves may be diverging strongly. In the two-counties trial the changing incidence was reflected by the mortality curve (Tabar *et al.*, 1985), and subsequent analyses (Tabar *et al.*, 1992) have demonstrated that it is strongly predictive of mortality. Much of the failure of our mortality reduction to increase between 7 and 10 years of follow-up is likely to be explained by delay in this aspect of performance. Although we have detected large numbers of *in situ* cancer, this is predicted to have less impact on mortality than reducing the rate of small node-positive cancers (Tabar *et al.*, 1992). One of the strengths of the present trial database is that high achievement in pathological classifications of nodal status within one city has permitted an analysis which identifies the problem more clearly than is possible using tumour size alone (UK Breast Cancer Detection Working Group, 1993).

When service screening was introduced into Edinburgh we took the opportunity of designing a protocol which would enable us to report on the interval cancer rates over 3 years in regularly screened women; these should be predictive of eventual rates in the service programme with its 3 yearly schedule. The optimum interval between examinations is unknown; a randomised trial comparing 1 yearly intervals with 3 yearly ones is currently in progress (N. Day, personal communication) but will not report for several years. Meanwhile, the best method of evaluating the inter-screening interval is to compare the proportional incidence of interval cases across the period (Tabar *et al.*, 1987). The proportions we report are based on small numbers but exceed those for the two-counties trial in both the second (42% compared with 29%) and third (67% compared with 45%) year (Tabar *et al.*, 1987). This indicates, firstly, that the interval could be too long and should certainly not exceed 3 years – for the population or for individual women. Secondly, it may be further evidence that screening as practised during the trial lacked sensitivity for detecting biologically important cancers.

In conclusion, this trial has provided modest but important contributions to the overall scientific evaluation of mammographic screening. These include further evidence that screening – at least for women over 50 years – can reduce mortality from breast cancer by around 20%. With current standards of mammography and higher attendance rates larger reductions may be achievable. The extension of results from scientific trials to routine health care can be problematic, but current reports from UK service screening are encouraging (Chamberlain *et al.*, 1993; Scottish Breast Screening Programme, 1993). There is a need to determine the best age for screening to commence. Since the benefits of screening take at least 4 years to emerge, the present service screening programme though targeted at women of 50 years and over will have little impact on mortality from the disease in women aged 50-54 years despite the fact that 20% of all deaths of British women in this age group are due to breast cancer (OPCS, 1989-93).

Finally, two warning messages emerge from this study: firstly, follow-up and recall facilities must ensure that minimal numbers of women wait for more than 3 years between their invitations to screening. Secondly, screening targets must focus on detection of cancers before nodal metastases develop rather than relying on favourable size. It follows that it is essential to the monitoring of the UK service screening programme that histological evidence of node status is available on all cases arising in the target population.

The Edinburgh randomised trial was funded by the Cancer Research Campaign and the Chief Scientists Office (CSO) of the Scottish Home and Health Department. We thank the project committee for their advice and encouragement over the fieldwork period. The members of the committee were: Dr F.E. Alexander, Dr T.J. Anderson, Dr M.M. Andrew, Professor J.J.K. Best, Dr C. Brough, Dr W. Forbes, Professor A.P.M. Forrest (Chairman), Dr R. Gruer, Dr A. Huggins, Dr L. Kinlen, Dr A.E. Kirkpatrick, Dr N.B. Loudon, W. Lutz and Dr U. MacLean, Dr M.M. Roberts with Mr John Duncan and Dr Ian Sutherland. We especially wish to record the contribution made to this trial by Dr Maureen Roberts, who was director of the Edinburgh breast screening clinic from 1979 to her death in 1989. We also thank Dr Michael Ashley Millar and Dr Boyd Moir for help in coordinating the funding and Dr Helen Stewart for her assistance in the collection of recent data and classification of review causes of death. We acknowledge the assistance of the staff at the General Register Office, Scotland, the Lothian Health Board Primary Health Care Division and general practitioners in Edinburgh as well as our colleagues in the Screening Clinic and the University Departments of Public Health Science, Pathology and Surgery. Additional secretarial support was provided by the Leverhulme trust.

References

ALEXANDER, F.E., ROBERTS, M.M., LUTZ, W. & HEPBURN, W. (1989). Randomisation by cluster and the problem of social class bias. *J. Epidermiol. Community Hlth*, **43,** 29–36.

ANDERSSON, I., ASPERGREN, K., JANZON, L. & 6 others (1988). Mammoraphic screening and mortality from breast cancer: the Malmo trial. *Br. Med. J.*, **297,** 943–948.

BERAL, V. (1993). Breast cancer: mammographic screening. *Lancet*, **341,** 1509–1510.

BRINKLEY, D., HAYBRITTLE, J.L. & ALDERSON, M.R. (1984). Death certification in cancer of the breast. *Br. Med. J.*, **289,** 465–467.

CHAMBERLAIN, J., MOSS, S.M., KIRKPATRICK, A.C., MITCHELL, M. & JOHNS, L. (1993). National Health Service breast screening programme results for 1991–2. *Br. Med. J.*, **307,** 353–356.

CHU, K.C., SMART, C.R. & TARONE, R.E. (1988). Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan clinical trial. *J. Natl Cancer Inst.*, **80,** 1125–1131.

EARLY BREAST CANCER TRIALISTS COLLABORATIVE GROUP (1992). Systematic treatment of early breast cancer by hormonal, cytotoxic or immunotherapy. *Lancet*, **i,** 1–17.

ELWOOD, J.M., COX, B. & RICHARDSON, A.K. (1993). The effectiveness of breast cancer screening in young women. *Curr. Clin. Trials*, **2,** 227–287.

FLETCHER, S.W., BLACK, W., HARRIS, R., RIMER, B.K. & SHAPIRO, S. (1993). Report of the international workshop on screening for breast cancer. *J. Natl Cancer Inst.* (in press).

FORREST, A.P.M. (CHAIRMAN) (1987). *Breast Cancer Screening*, report to the Health Ministers of England, Wales, Scotland and Northern Ireland by a working group. HMSO: London.

MILLER, A.B., BAINES, C.J., TO, T. & WALL, C. (1992). Canadian National Breast Screening Study: I. Breast cancer detection and death rates among women 40 to 49 years. *Can. Med. Assoc. J.*, **147,** 1459–1488.

MORRISON, A.S., BRISSON, J. & KHALID, N. (1988). Breast cancer incidence in the breast cancer demonstration project. *J. Natl Cancer Inst.*, **80,** 1540–1547.

NYSTROM, L., RUTQVIST, L.E., WALL, S. & others (1993). Breast cancer screening with mammography: an overview of the Swedish randomised trials. *Lancet*, **341,** 973–978.

OPCS MORTALITY STATISTICS: CAUSE. (1985). Series DH2 No. 11. HMSO: London.

OPCS CANCER STATISTICS: REGISTRATIONS. (1989–93). Series HBI Nos. 16–20 HMSO: London.

ROBERTS, M.M., ALEXANDER, F.E., ANDERSON, T.J. & 7 others (1984). The Edinburgh randomised trial of screening for breast cancer: description of method. *Br. J. Cancer*, **47,** 1–6.

ROBERTS, M.M., ALEXANDER, F.E., ANDERSON, T.J. & 9 others (1990). Edinburgh trial of screening for breast cancer: mortality at seven years. *Lancet*, **335,** 241–246.

SCOTTISH BREAST SCREENING PROGRAMME (1993). *Scottish Breast Screening Programme Report, 1993.* ISD Publications: Edinburgh.

SHAPIRO, S., VENET, W., STRAX, P. & ROESER, R. (1982). Ten-to-fourteen year effect of screening on breast cancer mortality. *J. Natl Cancer Inst.*, **69,** 349–355.

SHAPIRO, S. (1988). *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986.* Johns Hopkins University Press: Baltimore.

TABAR, L., FAGERBERG, C.J.G., GAD, A. & 8 others (1985). Reduction in mortality from breast cancer screening with mammography. *Lancet*, **i,** 829–832.

TABAR, L., FAGERBERG, C.J.G., GAD, A. & 2 others (1987). What is the optimal interval between mammographic screening examinations? An analysis based on the latest results of the Swedish two-county breast cancer screening trial. *Br. J. Cancer*, **55,** 547–551.

TABAR, L., FAGERBERG, G., DUFFY, S.W., DAY, N.E., GAD, A. & GRONTOFT, O. (1992). Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiol. Clin. N. Am.*, **30,** 187–210.

TOMATIS, L., AITIO, A., DAY, N.E. & 5 others (eds) (1990). *Cancer: Causes, Occurrence and Control.* IARC: Lyons.

UK BREAST CANCER DETECTION WORKING GROUP (1981). Trial of early detection of breast cancer: description of method. *Br. J. Cancer*, **44,** 618–623.

UK BREAST CANCER DETECTION WORKING GROUP (1993). Breast cancer mortality after 10 years in the UK trial of early detection of breast cancer. *Breast*, **2,** 13–20.

UK BREAST CANCER DETECTION WORKING GROUP (1991). Verification of the cause of death in the trial of early detection of breast cancer. *Br. J. Cancer*, **64,** 1151–1156.

VERBEEK, A.L.M., HENDRIKS, J.H.C.L., HOLLAND, R., MRAVUNAC, M. & STURMANS, F. (1985). Mammographic screening and breast cancer mortality: age-specific effects in the Nijmegen project, 1975–82. *Lancet*, **i,** 865–856.

WALD, N., FROST, C. & CUCKLE, H. (1991). Breast cancer screening: the current position. *Br. Med. J.*, **302,** 845.

WILLIAMS, D.A. (1982). Extra-binomial variation in logistic linear models. *Appl. Stat.* **31,** 144–148.