BMC Medical Research Methodology

BioMed Central

Research article

Imputation of a true endpoint from a surrogate: application to a cluster randomized controlled trial with partial information on the true endpoint

Richard M Nixon*[1], Stephen W Duffy[2] and Guy RK Fender[3]

Address: [1]MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, UK, [2]Cancer Research UK Department of Epidemiology, Mathematics and Statistics, Wolfson Institute of Preventive Medicine, Charterhouse Square, London, EC1M 6BQ, UK and [3]Taunton and Somerset Hospital, Department of Obstetrics and Gynaecology, Musgrove Park, Taunton, TA1 5DN, UK

Email: Richard M Nixon* - richard.nixon@mrc-bsu.cam.ac.uk; Stephen W Duffy - stephen.duffy@cancer.org.uk; Guy RK Fender - guy.fender@tst.nhs.uk

* Corresponding author

Published: 24 September 2003

BMC Medical Research Methodology 2003, **3**:17

Received: 10 April 2003
Accepted: 24 September 2003

This article is available from: http://www.biomedcentral.com/1471-2288/3/17

Abstract

Background: The Anglia Menorrhagia Education Study (AMES) is a randomized controlled trial testing the effectiveness of an education package applied to general practices. Binary data are available from two sources; general practitioner reported referrals to hospital, and referrals to hospital determined by independent audit of the general practices. The former may be regarded as a surrogate for the latter, which is regarded as the true endpoint. Data are only available for the true end point on a sub set of the practices, but there are surrogate data for almost all of the audited practices and for most of the remaining practices.

Methods: The aim of this paper was to estimate the treatment effect using data from every practice in the study. Where the true endpoint was not available, it was estimated by three approaches, a regression method, multiple imputation and a full likelihood model.

Results: Including the surrogate data in the analysis yielded an estimate of the treatment effect which was more precise than an estimate gained from using the true end point data alone.

Conclusions: The full likelihood method provides a new imputation tool at the disposal of trials with surrogate data.

Background

The Anglia Menorrhagia Education Study (AMES) [1,2] is a randomized controlled trial which tested the effectiveness of an "academic detailing" education package [3] in primary care and hospital gynaecology units to improve the management of women with menorrhagia (excessive menstrual bleeding). Here we are concerned with the first phase of this trial and only consider data from primary care.

The general practice was the unit of randomization and the primary outcome measure of interest was the proportion of referrals of women with menorrhagia to hospital. In this part of the trial, data were collected in two ways. Firstly the doctors in the practices in the study were asked

Table 1: Reported and audited outcome data

Trial phase Audited	Pre-intervention		Post-intervention	
	Intervention	Control	Intervention	Control
Patients seen	307	209	418	237
Referrals	56	39	80	63
Number of practices	27	25	27	25
Reported				
Patients seen	NA	NA	381	215
Referrals	NA	NA	93	92
Number of practices	NA	NA	40	36

to keep a record of consultations for menorrhagia, with outcome of consultation, on supplied data sheets. We refer to this as the reported data. Secondly, an audit of 52% of the practices was performed after the trial was over. This was performed in order to have an objective measure of referral which did not depend on a busy practitioner reporting. 52% of the practices was considered enough for sufficient power having seen the reported data. The reported data was only recorded for one year post-intervention, whereas one-year pre-intervention data was also available for the audited part of the trial. Total numbers of patients seen and patients referred for the reported and audits phase are given in Table 1. This paper is concerned with combining the reported and audited data from the primary care part of the trial. The reported data may be regarded as a surrogate for the audited data. There were 54 practices randomized to receive the education package and 46 to control. In the reported part of the study, 76 practices returned at least one data sheet (40 intervention, 36 control). The rest either returned no data sheets (5 practices) or no data sheets for menorrhagia (19 practices). It is conceivable that the surrogate reported endpoint data is missing because of some property of the practice which is related to the practice's likelihood of referring a patient, although this is unlikely to make a material difference to the results. No attempt is made to impute this missing reported data. 52 practices were chosen at random to be audited (27 intervention, 25 control). Of the practices audited, 50 also supplied reported data (26 intervention, 24 control). Hence partial data on the true endpoint and partial data on a surrogate are available.

In analysis, one might simply exclude those practices which do not have audited data. On the other hand, it is reasonable to suppose that some information, albeit less reliable, is contained in the reported data. Surrogate endpoints have been used in a variety of studies, notably in trials of cancer screening [4]. A criterion frequently used to assess the usefulness of a surrogate variable is the Prentice criterion [5], which stipulates that the effect of the treatment on the true endpoint is entirely attributable to its effect on the surrogate. Begg and Leung [6] point out that even if this holds, the absolute magnitude of the effect on the true endpoint may be different from the magnitude of the effect on the surrogate. Begg and Leung also contend that it is more important that the surrogate be strongly correlated with the true endpoint. It is therefore desirable to estimate the effect of the surrogate on the true endpoint, even if only surrogate information is available. In our case, we have true audited endpoint data on 52 of the 100 study practices. Of these 50 out of 52 (96%) also have surrogate reported endpoint data. Of the remaining 48 practices, 26 (54%) provided surrogate data only and 22 (46%) provided no data at all. From this information, it is possible in principle to estimate the relationship between the surrogate and the true outcome, and therefore the trial results for all 78 practices providing surrogate or true end point data or both. In this paper we are using the surrogate variable to strengthen inference about the true endpoint. In this case the surrogate variable is known as an auxiliary variable [7].

Three approaches are considered. A regression method, multiple imputation and a full likelihood model. The regression method is a three stage process. Firstly, the observed audited data is modelled as a function of the corresponding reported data and general practice characteristics. Secondly, the missing audited data is generated using the parameter estimates from this modelling. Thirdly, a random effects model is fitted to assess the effectiveness of the intervention. This model also includes a term that denotes whether the true endpoint was observed or estimated. Multiple imputation generates several realisations of the missing audited data, given the observed data. Each of these imputed data sets is then used to generate an estimate of the effectiveness of the intervention. Finally each of these estimates is combined to give an overall

Table 2: General practice characteristics

Practice characteristic	All audited patients				All patients with reported data			
	Intervention		Control		Intervention		Control	
Mean list size	6974		5167		6965		5314	
Fund-holding	7/27	26%	4/25	16%	11/40	28%	6/36	17%
Has branch surgeries	15/27	56%	16/25	64%	17/40	43%	24/36	67%
Rural	10/27	37%	7/25	28%	16/36	44%	9/36	25%
Has drug dispensing facilities[1]	0.34		0.42		0.32		0.44	
Male partners[1]	0.63		0.77		0.67		0.76	
Has trainees	15/27	56%	9/25	36%	15/40	38%	9/36	25%
Partners on obstetric list[1]	0.92		1.00		0.89		0.99	

1 = characteristic is measured as a mean proportion

estimate of the true outcome effect. The full likelihood model generates an imputation of the missing audited data from the reported and audited data, and performs the randomized trial comparison simultaneously. In all these approaches, we are assuming the audit data is missing at random (MAR) [8] i.e. the missing audit data mechanism is only dependent on observed reported data (and also observed practice characteristics in the regression method). The exception to this is in the first of two multiple imputation approaches used. The missing data is generated solely from the observed audit outcome data. This is a missing completely at random (MCAR) mechanism as the reason an audit outcome is missing is assumed not to depend on any other observed or missing values. MAR and MCAR both assume the reasons for missing data do not depend on knowing unobserved data. A final type of missing mechanism, not assumed in this paper, is not missing at random (NMAR). Here the reason for missing audit data depends on unrecorded missing values. Data that are MCAR and MAR are sometimes referred to as "ignorable" because estimation of the model parameters is valid even if one does not estimate the parameters of the missing data mechanism. However, data this is NMAR is referred to as "non-ignorable" because estimates of model parameters are invalid if one does not estimate the parameters of the missing data mechanism. Date available are summarised in Table 1. General practice characteristics are shown in Table 2 by trial arm.

The aim of this paper is to estimate the treatment effect using data from every practice in the study. Where the true endpoint is not available, it is estimated via a surrogate by three approaches, a regression method, multiple imputation and a full likelihood model.

Methods

Since our endpoint was referral of individual patients, but the unit of randomization was general practice, all models assessing the treatment effect incorporated a random effects component for practice, to take account of this cluster randomization [9]. Logistic regression is used in all the analysis. If r, the number of referrals is 0; or if $r = n$, the number of patients seen, this causes problems with some of the methods used. When this occures r is replaced by $r + 0.5$ and n is replaced by $n + 1$ [10]. For consistency all analysis are performed on the same data set, regardless of whether this change is necessary for a particular analysis.

Regression models

For the 26 practices which were not audited, but for which we had reported data, our aim was to predict what audit data (pre and post intervention numbers of patients seen and referred) would have come from these practices had they been audited. We used the post intervention reported data to predict both the pre and post intervention audited data. We also included practice characteristics in this prediction. Log linear regression models were fitted, where the audited data is a function of the reported data and the practice characteristics. Then estimated parameters from these models were used to estimate the missing values from the audited data. Finally, the overall effect of intervention was estimated by a random effects logistic regression model, where an extra random effect was included to add extra variance from the observations that were estimated and not observed.

Firstly, we fitted log linear regression models of audit data on reported data and practice characteristics. Randomization group status was included in the practice characteristics vector in the pre-intervention regression, as reported data, the crucial independent variable, is only observed after the intervention, and the relationship between reported behaviour after intervention and true behaviour before intervention may be modified by the effect of the intervention (e.g. a reduction in referral rates after intervention in the groups receiving the intervention). On the other hand, in the post-intervention regression, the reported and audited data are observed post-

intervention, so the effect of intervention is already included. The following models are fitted for the 50 practices that were audited and which returned at least one reported data form:

$$\begin{aligned}
\log\left(n^b\right) &= \alpha_1 + \beta_1^T P_1 + \gamma_1 \log\left(n^r\right) \\
\log\left(n^a\right) &= \alpha_2 + \beta_2^T P_2 + \gamma_2 \log\left(n^r\right) \\
\log\left(r^b\right) &= \alpha_3 + \beta_3^T P_1 + \gamma_3 \log\left(r^r\right) \\
\log\left(r^a\right) &= \alpha_4 + \beta_4^T P_2 + \gamma_4 \log\left(r^r\right) \qquad (1)
\end{aligned}$$

n^b and n^a denote the number of women presenting with menorrhagia before and after intervention from the audited data respectively; n^r denotes the corresponding number from the post-intervention reported data. r^b, r^a and r^r are the corresponding number of referrals from the pre- and post-intervention audited and post-intervention reported data respectively. p_2 is the vector of the eight practice characteristics given in table 2, and p_1 this same vector, but also including the randomization group of the practice.

The values of α, β and γ are used to generate fitted values for the 76 practices which have reported data. A full data set can now be constructed for all 78 practices with any data at all. For the 52 practices with observed audited data, this is used, and the fitted values are ignored. Plots of observed data verses fitted values for the 50 practices that supplied both audited and reported data are shown in Figure 1. This is to gauge visually how well the reported data predicts the audited. For the 26 practices which only had reported data, the fitted audit values are used. This data is then used in fitting the following random effects model:

$$\begin{aligned}
r_{ijkl} &\sim \mathrm{Bi}\left(n_{ijkl}, \pi_{ijkl}\right) \\
\log\left(\frac{\pi_{ijkl}}{1-\pi_{ijkl}}\right) &= \alpha + \beta R_{jk} + \gamma_i\left(T_k - 0.5\right) + \delta_i E_l + \varepsilon_{ijkl} \\
\gamma_i &\underset{\sim}{i.i.d} N\left(\mu_1, \sigma_1^2\right) \\
\delta_i &\underset{\sim}{i.i.d} N\left(0, \sigma_2^2\right) \qquad (2)
\end{aligned}$$

Where n_{ijkl} and r_{ijkl} are the number of women presenting with menorrhagia, and the number of women referred in practice i, from intervention group j (0 = control, 1 = intervention), in study period k (0 = pre-, 1 = post-intervention). This comes from observed audited data where available (l = 0), and fitted audited data where it is missing (l = 1). π_{ijkl} denotes the true underlying probability of being referred. R, T and E are dummy variables: $R_{00} = R_{01} = R_{10} = 0$ (control group and intervention group pre-intervention), R_{11} = 1 (intervention group post-intervention), T_0 = 0 (pre intervention), T_1 = 1 (post intervention) and E_0 = 0 (observed), E_1 = 1 (estimated). β is used to denote the log odds ratio of being referred in an intervention practice post intervention compared to a control practice or intervention practice pre intervention. In this model we allow a variation in trend for each practice γ_i, around an average trend for all practices μ_1. There is a common intercept for each practice, within trial arm, at the point (T_k - 0.5). δ_i is a random effect that is "switched off" for the practices that have observed audited values and "switched on" for the practices that use estimated audited values. In this way extra variability is allowed in the model for the practices that have estimated audit information.

Multiple imputation

Methodology

Multiple imputation using auxiliary variables can be used to strengthen the true endpoint [11]. In our case we use an approximate Bayesian bootstrap method. Each practice's audited data is regarded as a single data point. The basic method is to take bootstrap re-samples of the known audited data and from these, take smaller bootstrap samples to simulate the missing data. Underlying this is the theory that we are sampling from a scaled multinomial distribution as an approximation to a Dirichlet posterior distribution [12]. Thus the data has to be expressible as a realisation of a discrete categorical variable, which in this case holds, albeit with a large number of possible realisations.

More formally suppose we have a vector of discrete data Y, that contains observed values Y_{obs} and missing values Y_{mis}. Y can take values $d_1 \ldots d_K$ with probabilities $\theta = (\theta_1, \ldots, \theta_K)$ respectively. Rubin [13] defines a Bayesian bootstrap implementation. A Dirichlet prior distribution is defined for θ, from a non-informative prior. One realisation, θ*, of θ is drawn from its posterior. Finally the components of Y_{mis} are independently drawn from $d_1 \ldots d_K$, such that the P (drawing d_k) = θ_k^* k = 1...K. This gives one imputation of the complete data Y. The process is repeated M times to get M multiple imputations.

The Bayesian bootstrap imputation is complicated to implement as it requires sampling from a Dirichlet distribution, followed by taking a weighted sample from the possible values the components Y can take. There is a simple approximation to the Bayesian bootstrap method, also defined by Rubin [13] which is easier to compute in practice.

Suppose Y_{obs} and Y_{mis} is of length n_0 is of length n_1. The approximate Bayesian bootstrap imputation is as follows:

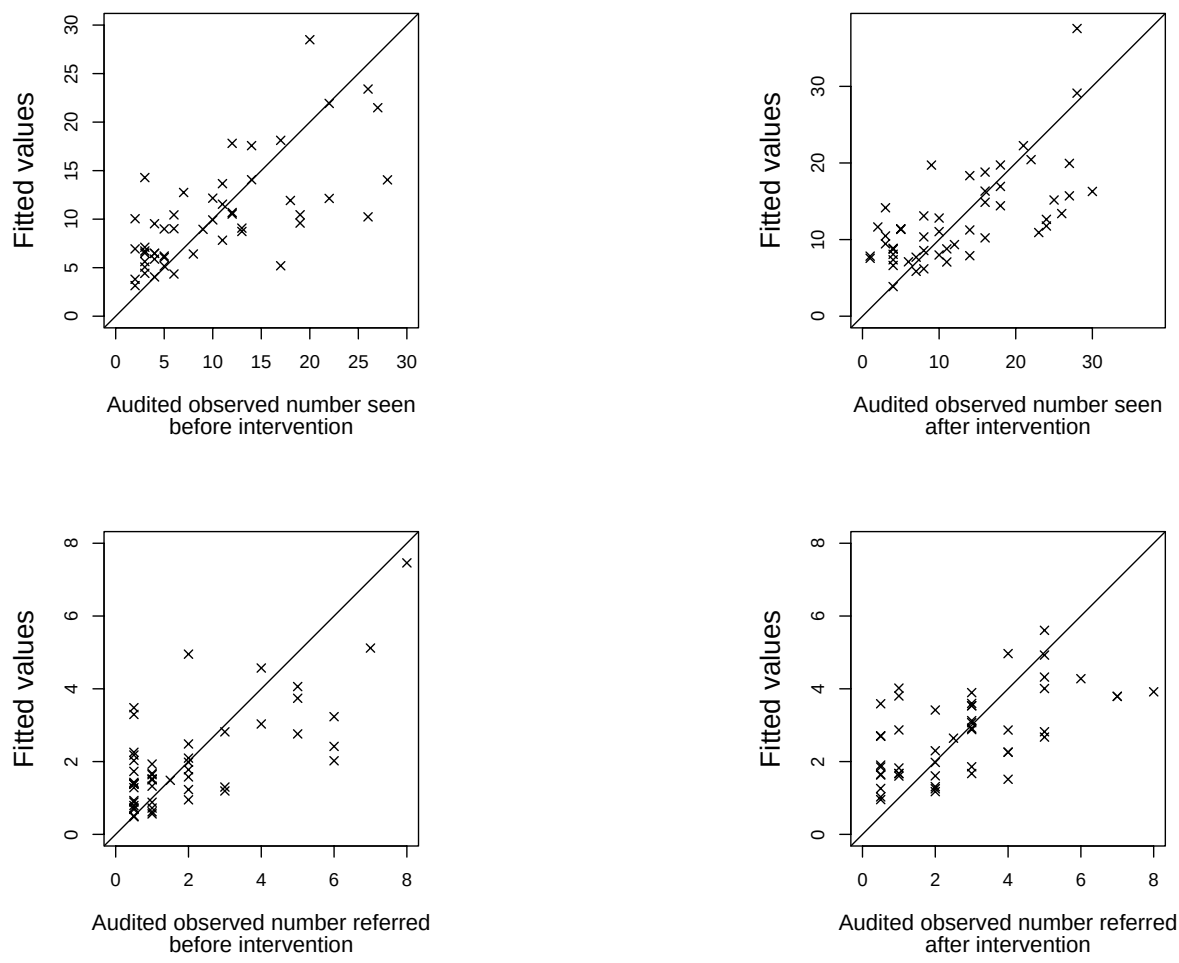

Figure 1
Plots of observed versus fitted values for the 50 practices that supplied audited and reported data. The observed values correspond to the audited information recorded, the fitted values correspond to the audited information that is predicted from the reported data via the regression model. Lines with a zero intercept and a gradient of one are plotted to gauge agreement between observed and fitted values.

• Draw n_1 components, with replacement, from Y_{obs}. Call this vector Y^*_{obs} .

• Draw n_0 components, with replacement from Y^*_{obs} . This sample is the imputed Y_{mis}.

In this way the approximate Bayesian bootstrap method draws θ from a scaled multinomial distribution rather then a Dirichlet posterior as in the Bayesian bootstrap case.

Suppose we wish to estimate β from the data. *M* different point estimates of β, and its variance will be estimated from each of the imputed data sets, and we call these $\hat{\beta}^{(1)},...\hat{\beta}^{(M)}$. Rubin [14] gives the following rule for combining these estimates into a single estimate. The combined point estimate is the average of the *M* point estimates from the imputed data:

$$\bar{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)} \qquad (3)$$

Variance of $\bar{\beta}$ comes from two sources. The within-imputation variance, which is the average of the variances of the $\hat{\beta}^{(m)}$:

$$W = \frac{1}{M}\sum_{m=1}^{M}\mathrm{var}\left(\hat{\beta}^{(m)}\right) \qquad (4)$$

and the between-imputation variance, which is the variance of the estimates of $\hat{\beta}^{(m)}$:

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}\right)^2 \qquad (5)$$

The total variance is defined as:

$$T = W + (1 + M^{-1})B \qquad (6)$$

Inferences about β can be gained from the approximation:

$$\frac{\beta - \bar{\beta}}{\sqrt{T}} \sim t_v \qquad (7)$$

Where the degrees of freedom of the t distribution is given by:

$$v = (M-1)\left[1 + \frac{W}{\left(1 + M^{-1}\right)B}\right]^2 \qquad (8)$$

So β is estimated by $\bar{\beta}$ and the CI given by:

$$\bar{\beta} \pm t_{v,1-\frac{\alpha}{2}}\sqrt{T} \qquad (9)$$

where $t_{v,1-\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ point of the t-distribution on v degrees of freedom. Further details of the basics of multiple imputation are available from Schafer [15].

Application to the AMES data set

Let a be the fraction of missing information for a scalar estimator. Rubin [14] calculates that the relative efficiency (on the variance scale) of a point estimate based on M imputations compared to one based on an infinite number of imputations is approximately:

$$\frac{1}{1 + \frac{a}{M}} \qquad (10)$$

In this case $a = 26/78 = 1/3$ (52 practices have audited data and we impute for the 26 that have this data missing) so if we set $M = 5$ the s.e. of the estimate will be $\sqrt{(1 + 1/15)} = 1.033$ times as large as the estimate with $M \to \infty$.

We are only interested in imputing the missing audited data, as ultimately this is considered the most accurate, and the pre-intervention data can be used in the modelling. The missing audit data always comes in groups of four for each practice: the number of women presenting with menorrhagia and the number of referrals both pre and post intervention. The theory outlined above is for imputation of missing data in a vector. We identify the audited data for each practice (i.e. row of the data) with an element of a vector Y. That is to say, each element of Y contains the audited data for one practice. In this way data is always imputed per practice and not individually for each field.

The approximate Bayesian bootstrap imputation was then performed on the data. A random sample of 52 rows was taken with replacement from the 52 rows of complete data. From this a random sample of 26 rows was taken with replacement. This, along with the original 52 complete rows forms an imputed data set. This process was independently repeated five times, and these five data sets are each analysed.

As with the analysis of the audit data before, we wish to get an estimate of the odds of being referred in the intervention group compared to the control group. We fit the model:

$$\begin{aligned} r_{ijkl} &\sim \mathrm{Bi}\left(n_{ijkl}, \pi_{ijkl}\right) \\ \log\left(\frac{\pi_{ijkl}}{1-\pi_{ijkl}}\right) &= \alpha + \beta R_{jk} + \gamma_i\left(T_k - 0.5\right) \\ \gamma_i &\overset{i.i.d.}{\sim} N\left(\mu, \sigma^2\right) \end{aligned} \qquad (11)$$

Where the variable definitions are the same as those used in equation 2.

This imputation assumes that the missing audit data is missing completely at random (MCAR) [8] as all the missing data comes from the same distribution and pays no attention to the reported data when imputing the missing data. As the number of patients reported to have been seen and referred to hospital may be informative for the audited values, then it is desirable that the imputation

process includes the reported data in the estimation of the missing audited data. The missing audit data is now assumed to be missing at random. To do this, the data set was stratified by reporting behaviour. Six strata were defined by the total number of patients reported to have been seen (either ≤ 6 or ≥ 7), and the proportion of patients reported to have been referred, ([0,0.15), [0.15,0.4), [0.4,1.0]). These categories were chosen as the median number of patients reported to have been seen was 6.5 and the 33^{rd} and 66^{th} percentiles of the proportion of patients reported to have been referred were 0.15 and 0.4. Within each strata the missing data were then imputed from the observed data.

52 practices have observed audited data. In the stratified imputation only 50 of these can be used to sample from, as two of these practices have no reported data on which to stratify.

Full likelihood model

The previous two methods used a two-stage procedure where firstly missing data was imputed and then the treatment effect estimated. Here a method is proposed that performs both these stages simultaneously. Consider the following model:

$$
\begin{aligned}
n_{ij}^{a} &\sim \mathrm{Po}\left(\phi^{a}\right) \\
r_{ij}^{a} &\sim \mathrm{Bi}\left(n_{ij}^{a}, p_{ij}^{a}\right) \\
\mathrm{logit}\left(p_{ij}^{a}\right) &= \alpha_1 + \beta_1 R_j + \gamma_i^{a} \\
\gamma_i^{a} &\sim \mathrm{N}\left(0, \tau^{a}\right) \\
\\
n_{ij}^{r} &\sim \mathrm{Po}\left(\phi^{r}\right) \\
r_{ij}^{r} &\sim \mathrm{Bi}\left(n_{ij}^{r}, p_{ij}^{r}\right) \\
\mathrm{logit}\left(p_{ij}^{r}\right) &\sim \mathrm{N}\left(\mu_{ij}^{r}, \tau^{\tau}\right) \\
\mu_{ij}^{r} &= \alpha_2 + \beta_2 \mathrm{logit}\left(p_{ij}^{a}\right) \qquad (12)
\end{aligned}
$$

Note that this model uses only the post-intervention data and does not use the practice characteristics data. Here n_{ij}^{a} and r_{ij}^{a} denote the number of women presenting with menorrhagia and the number of referrals in practice i in treatment group j from the audited data. n_{ij}^{r} and r_{ij}^{r} are the corresponding numbers from the reported data. p_{ij}^{a} is assumed to be the true underlying audited probability of referral, and p_{ij}^{r} the true underlying surrogate (i.e. reported) probability of referral. A Directed acyclic graph of this model is given in Figure 2. The model was fitted using MCMC sampling [16].

Neither the audited data nor the reported data is complete. Of the 78 practices included in this model, 76 have reported data and 52 have audited data (50 have both). Because of the nature of the MCMC sampler used in the model fitting, at each iteration the observed values and the current imputed values of n_{ij}^{a} and n_{ij}^{r} were used to estimate ϕ^{a} and ϕ^{r} respectively; in turn ϕ^{a} and ϕ^{r} then impute another set of missing values of n_{ij}^{a} and n_{ij}^{r}.

For each practice the logit true audited probability of referral, logit(p_{ij}^{a}), is modelled as a linear function of treatment group. Thus β_1 is an estimate of the log odds of referral in the intervention group compared to the control. Each practice as allowed to have a different underlying probability of referral via the random effect γ_i, thus making adjustments for the cluster randomized nature of the design.

The reported data was considered to be a surrogate for the audited data. In this model the logit surrogate probability of referral, logit(p_{ij}^{r}), was assumed to be a linear function of the logit audited probability of referral logit(p_{ij}^{a}). As missing audit data is dependent on reported data alone then it is assumed to be missing at random.

In MCMC sampling, at every iteration an estimate of every parameter is obtained. This means that missing data imputation and the randomized trial comparison were performed simultaneously and not in a two stage process.

Model fitting

The regression models given in equation 1, that are used to generate missing values, were fitted using Splus [17]. The random effects model given in equation 2, to estimate the odds of referral in an intervention practice compared to a control, was fitted using BUGS [18]. The approximate Bayesian bootstrap imputed data sets were generated using Splus; the separate log odds ratios for each imputed data set calculated from equation 11 were fitted using BUGS; and the overall multiple imputed estimate was generated by code written in Splus. The full likelihood model was fitted using BUGS.

Where BUGS was used to estimate parameters, prior distributions that are locally almost uniform were chosen, with variances at least two orders of magnitude larger than

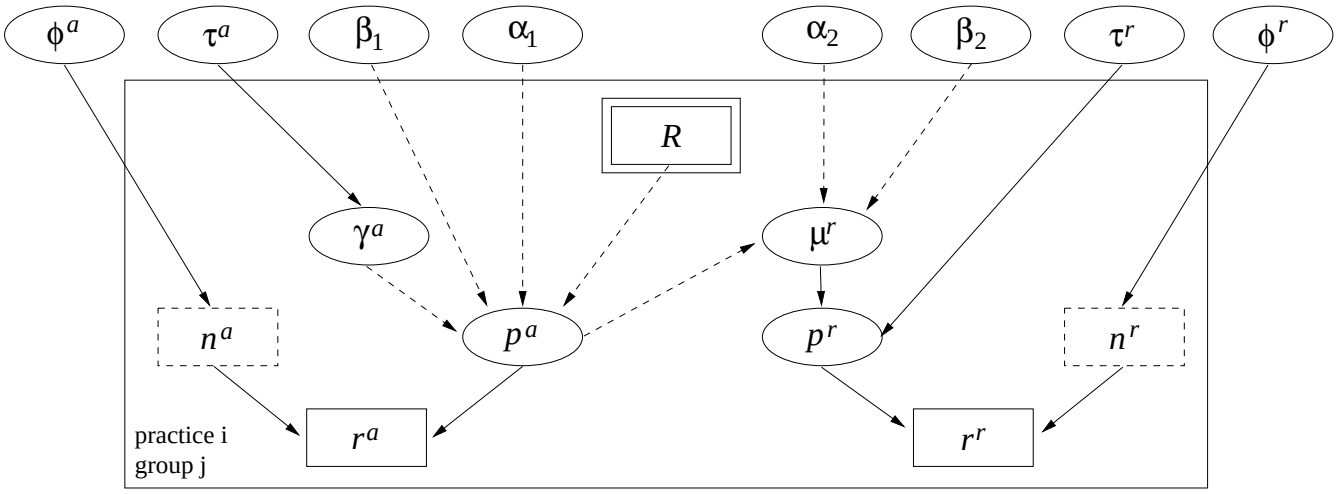

Figure 2
Directed Acyclic Graph of the full likelihood model. In the graph circles represent unknown parameters and rectangles represent observed data. Dashed arrows represent deterministic dependence and solid arrows represent stochastic dependence. A dashed rectangle represents data that is partially observed, and is imputed (so a parameter) for missing values.

Table 3: Odds of being referred in an education practice compared to a control: comparison of the various modelling strategies used.

Method set	Point estimate	CI	s.e. (log OR)
Audited data only	0.73	(0.47,1.08)	0.212
Regression	0.68	(0.42,1.01)	0.218
Unstratified imputation	0.74	(0.45,1.02)	0.203
Stratified imputation	0.75	(0.45,1.05)	0.212
Full likelihood	0.68	(0.44,0.91)	0.188

Table 4: The correlation between the audited and reported data for the number of referrals *r*, the total number of patients seen *n* and the proportion of referrals *r*/*n*, for all the post-intervention data where both are available.

r	*n*	*r*/*n*
0.17	0.36	0.30

the posterior variances of the corresponding nodes. These priors are considered to be non informative. The model fitting for the full likelihood model was achieved with the BUGS code in the Appendix. Convergence was assessed by the methods of Geweke [19], Raftery & Lewis [20], and Heidelberger & Welch [21], using the BOA package [22].

Results

Table 3 shows the results of the three methods, together with the results obtained when using a random effects logistic regression model on the audited data alone. All methods showed a reduction in the odds of referral of around 30%. The greatest precision was achieved by the full likelihood model.

Table 4 shows the correlation between the audited and reported data for the number of referrals, the total number of patients seen and the proportion of referrals for all post-intervention data where both are available. These correlations are quite low, which helps to explain why the gains in precision of the estimated treatment effect are moderate when the reported data are used.

Discussion

These results show reasonable agreement with regard to the point estimate. The educational package reduced the

proportion of women who are referred to hospital by around 30%. Some of this benefit may be artificial, due to increased diagnostic activity in the intervention group.

These results differ from those previously reported [2,9], as when performing the analysis using audited data alone, women were excluded who had post coital bleeding, pelvic mass, or bleeding between periods. As these exclusion criteria could not be applied to the surrogate reported data, they were not applied in the analysis reported here.

In all the modelling strategies used we have attempted to impute missing data from surrogate data and assess the effect of intervention. Each strategy has used different methods for imputing data, and for adjusting the variance of the outcome measure to allow for the fact that this data is estimated rather than observed.

The regression models attempt to account for extra variation, caused by using imputed values, by adding a random effect in the regression model which estimated the outcome measure. However, a weakness in this model is that the estimated values are artificially too good.

The fitted values will all lie on hyper-planes defined by the estimated parameters from model 1, whereas the observed values used when these are available will all lie around these planes, but will never lie exactly on them. The regression models used to estimate the missing values are therefore giving the exact values that one would expect and do not allow for random variation in the realised values. Extra variation is allowed in the model for these values by inclusion of an additional random effect. However, there is an element of a "self fulfilling prophecy" where the regression model 2 that estimates the outcome measure is based on data that will fit the model better at the estimated points.

A further problem with this method is that it is possible in general for r to be larger than n. This did not happen in this case as $\hat{\gamma}_1 > \hat{\gamma}_3$ and $\hat{\gamma}_2 > \hat{\gamma}_4$. An alternative modelling strategy to protect against this could be to estimate the number of events n and the probability of a positive outcome π from the surrogates. Then estimate the number of positive outcomes as $n\pi$.

The imputation models do not have these problems, as the missing audited data is imputed from the observed audited data. The stratified method is to be preferred as this generates data which is more likely to have occurred for the practice that the missing data is being imputed for. The results from the imputation methods give estimates of the effect of intervention and s.e. of the log odds ratio in between the other two methods. It should be noted that the formulas used to estimate this standard error have been shown to be inconsistent in certain settings [23]. The stratification goes some way towards imputing missing values which are appropriate for the practice, but it is still quite a blunt tool. An alternative method which could be considered is derived by Shafer [12]. Multiple imputation of multivariate categorical data under log linear models could be used. This method is based on the EM algorithm, where the likelihood function used for imputing the missing values can include a number of covariates. In this case the reported data, along with the practice characteristics could be used in the imputation process. An elegant application of the EM algorithm in estimation of missing data is given by Longford *et al* [24].

Table 5: Odds of being referred in an education practice compared to a control: results from the individual multiple imputations.

	Imputation	Point estimate of OR	s.e.
Unstratified	1	0.70	0.116
	2	0.73	0.115
	3	0.67	0.108
	4	0.85	0.123
	5	0.72	0.116
Stratified	1	0.84	0.136
	2	0.67	0.108
	3	0.71	0.114
	4	0.72	0.117
	5	0.82	0.137

The full likelihood model has the desirable property of performing the imputation and the randomized trial comparison simultaneously. Despite not making use of pre-intervention information, this model achieves the lowest standard error of the log odds ratio of all the models considered. The validity of the point estimate is unlikely to be impaired by the absence of pre-intervention data as the audited pre-intervention probabilities of referral were similar in the intervention and control groups. This model could be improved in principle by including pre-intervention data and practice characteristics. This was tried and imposed too heavy a burden on the estimation algorithm.

The standard error of the log odds ratio obtained from a random effects logistic regression on the audited data alone was 0.212. This was improved upon by the methods here which estimate the missing audited data, with the exception of the regression method, which was conservative, probably due to too much extra variation being add by the random effect for imputed values. These improvements are due to the added information from the auxiliary reported variable. The choice of parametric assumptions

used in the generation of missing values would also influence this gain in precision.

The reported data is strongly related to the audited data. The relationship of the logit reported probability with the logit audited probability is

$$\text{logit}(P^r) = 5.44 + 1.09\ \text{logit}(P^a) \quad (13)$$

The 95% credible interval of the estimate of 1.09 is (0.17,2.02), indicating a significant (p = 0.02) dependency of surrogate on true endpoint. Thus, while this surrogate is unlikely to satisfy Prentice's criteria [5], it does satisfy Begg and Leung's [6].

Conclusion

Using reported data as a surrogate for audited in the full likelihood model gives a point estimate that is accurate, and improves the precision of the estimate from that yielded using audited data alone. Regression type approaches and the Bayesian bootstrap imputation technique have already been used in other studies. The full likelihood approach provides an additional possible strategy in the case where only partial information is available on the true endpoint.

Competing interests

None declared.

Authors' contributions

RN developed the models, performed all the analysis and drafted the manuscript. SD aided in the development of the models and production of the final manuscript. GF designed and coordinated the AMES study.

Appendix: BUGS code for full likelihood model

```
model{

for(i in 1 : N){

  ref.a [i] ~ dbin(p.a [i], tot.a [i])

  tot.a [i] ~ dpois(phi.a)

  logit(p.a [i]) <- alpha1 + beta1 * treat [i] + gamma.a [i]

  gamma.a [i] ~ dnorm(0, tau.a)

  ref.rep [i] ~ dbin(p.rep [i], tot, rep [i])

  tot.rep [i] ~ dpois(phi.rep)

  lp.a [i] <- logit(p.a [i])

  logit(p.rep [i]) <- alpha2 + beta2 * (lp.a [i]- Ip.a.bar)

  }

Ip.a.bar <- mean(lp.a[])

tau.a <- l/(s.a*s.a)

s.a <- exp(ls.a)

#PRIORS

phi.a ~ dnorm(0,1.0E-6) I(0,)

phi.rep ~ dnorm(0,1.0E-6) I(0,)

ls.a ~ dunif(-6,6)

a ~ dnorm(0.0,1.0E-6)

b ~ dnorm(0.0,1.0E-6)

c ~ dnorm(0.0,1.0E-6)

d ~ dnorm(0.0,1.0E-6)

#EXTRA VARIABLES

exp.b <- exp(b)

}
```

References

1. Fender GRK, Prentice A, Gorst T, Nixon RM, Duffy SW, Day NE and Smith SK: **The anglia menorrhagia education study: A randomised controlled trial of an educational package on the management of menorrhagia in primary care.** *British Medical Journal* 1999, **318:**1246-1250.
2. Fender GRK, Prentice A, Nixon RM, Gorst T, Duffy SW, Day NE and Smith SK: **Management of menorrhagia: An audit of practices involved in the Anglia Menorrhagia Education Study (AMES).** *British Medical Journal* 2001, **322:**523-524.
3. Soumerai SB and Avorn J: **Principles of educational outreach (academic detailing) to improve clinical decision-making.** *Journal of the American Medical Association* 1990, **263:**549-556.
4. Day NE and Duffy SW: **Trial design based on surrogate endpoints – application to comparison of different screening frequencies.** *Journal of the Royal Statistical Society A* 1996, **159:**40-60.
5. Prentice RL: **Surrogate endpoints in clinical trials: definition and operational criteria.** *Statistics in Medicine* 1989, **8:**431-440.
6. Begg CB and Leung DHY: **On the use of surrogate end points in randomized trials.** *Journal of the Royal Statistical Society A* 2000, **163:**15-24.
7. fFeming TR, Prentice RL, Pepe MS and Glidden D: **Surrogate and auxiliary end-points in clinical-trials, with potential applications in cancer and aids research.** *Statistics in Medicine* 1994, **13:**955-968.
8. Little RA and Rubin DB: *Statistical analysis with missing data New York: Wiley*; 1987.
9. Nixon RM, Duffy SW, Fender GRK, Day NE and Prevost TC: **Randomisation at the level of General Practice : Use of pre intervention data and random effects models.** *Statistics in Medicine* 2001, **20:**1727-1738.
10. Cox DR and Snell EJ: *Analysis of binary data Chapman and Hall, London*; 1989.

11. Faucett CL, Schenker N and Taylor JMG: **Survival analysis using auxiliary variables via multiple imputation, with application to aids clinical trial data.** *Biometrics* 2002, **58:**37-47.
12. Schafer JL: *Analysis of incomplete multivariate data. Monographs on statistics and applied probability Chapman & Hall, London*; 1997.
13. Rubin DB: **The bayesian bootstrap.** *Annals of statistics* 1981, **9:**130-134.
14. Rubin DB: *Multiple imputation for nonresponse surveys J Wiley & son, New York*; 1987.
15. Schafer JL: **Multiple imputation: a primer.** *Statistical Methods in Medical Research* 1999, **8:**3-15.
16. Gelfand AE and Smith AFM: **Sampling-based approaches to calculating marginal densities.** *Journal of the American Statistical Association* 1990, **85:**398-409.
17. **Data analysis products division.** *S-Plus user's guide MathSoft*; 1997.
18. Gilks WR, Thomas A and Spiegelhalter DJ: **A language and program for complex bayesian modelling.** *The Statistician* 1994, **43:**169-177.
19. Geweke J: **Evaluating the accuracy of sampling-based approaches to calculating posterior moments.** In *Bayesian Statistics 4* Edited by: *Bernardo LM, Berger JO, Dawid AP, Smith AFM. Oxford University Press*; 1992.
20. Raftery AL and Lewis S: **How many iterations in the Gibbs sampler?** In *Bayesian Statistics 4* Edited by: *Bernardo LM, Berger JO, Dawid AP, Smith AFM. Oxford University Press*; 1992:763-774.
21. Heidelberger P and Welch PD: **Simulation run length control in the presence of an initial transient.** *Operations Research* 1983, **31:**1109-1144.
22. Smith B: **Bayesian output analysis program.** [http://www.public-health.uiowa.edu/boa].
23. Robins JM and Wang N: **Inference for imputation estimators.** *Biometrika* 2000, **87:**113-124.
24. Longford NT, Ely M, Hardy R and Wadsworth MEJ: **Handling missing data in diaries of alcohol consumption.** *Journal of the Royal Statistical Society: Series A* 2000, **163:**381-402.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/3/17/prepub