IJE vol.34 no.1 © International Epidemiological Association 2004; all rights reserved.
Advance Access publication 27 August 2004

International Journal of Epidemiology 2005;**34**:89–99
doi:10.1093/ije/dyh297

Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes

Angela M Wood,[1] Ian R White,[1] Melvyn Hillsdon[2] and James Carpenter[3]

Accepted 30 June 2004

Background Longitudinal studies almost always have some individuals with missing outcomes. Inappropriate handling of the missing data in the analysis can result in misleading conclusions. Here we review a wide range of methods to handle missing outcomes in single and repeated measures data and discuss which methods are most appropriate.

Methods Using data from a randomized controlled trial to compare two interventions for increasing physical activity, we compare complete-case analysis; *ad hoc* imputation techniques such as last observation carried forward and worst-case; model-based imputation; longitudinal models with random effects; and recently proposed joint models for repeated measures data and non-ignorable dropout.

Results Estimated intervention effects from *ad hoc* imputation methods vary widely. Standard multiple imputation and longitudinal modelling agree closely, as they should. Modifying the modelling method to allow for non-ignorable dropout had little effect on estimated intervention effects, but imputing using a common imputation model in both groups gave more conservative results.

Conclusions Results from *ad hoc* imputation methods should be avoided in favour of methods with more plausible assumptions although they may be computationally more complex. Although standard multiple imputation methods and longitudinal modelling methods are equivalent for estimating the treatment effect, the two approaches suggest different ways of relaxing the assumptions, and the choice between them depends on contextual knowledge.

Keywords Missing data, dropout, last observation carried forward, imputation, longitudinal data, random effects, non-ignorable

When individuals are followed over time in a longitudinal study or a randomized trial, it is almost inevitable that some outcomes will be unobserved. To draw inferences from such data it is almost always necessary to make some assumptions. A popular method of analysis restricts attention to individuals for whom the outcome of interest is observed. This makes the analysis simple (complete-case analysis), but the implicit assumption is very restrictive—namely, that the excluded individuals do not differ systematically in any way from the included individuals—and the analysis may be inefficient. It is therefore important to consider other methods of analysis.

Although our application is a randomized trial, the issues surrounding the handling of missing outcome data in observational studies are very similar. The only feature specific to randomized trials is the intention-to-treat principle, which states that all randomized individuals should be included in the analysis, irrespective of subsequent dropout or treatment changes;[1] however, it is also often desirable to include all recruited individuals in the analysis of observational studies. We do not focus on the different issues which arise with missing values in exposures and confounders in observational studies[2]

[1] MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK.

[2] Department of Epidemiology and Public Health, University College London, 1–19 Torrington Place, London WC1E 6BT, UK.

[3] Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Correspondence: Dr Angela Wood, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK. E-mail: angela.wood @mrc-bsu.cam.ac.uk

or with missing values in baseline covariates.[3] Missing baseline data in randomized controlled trials usually do not lead to bias but can reduce precision.[3]

A recent survey of the way randomized trials published in general medical journals handle missing data showed that 41 out of 63 trials used a complete-case analysis as the primary analysis.[4] Fifteen trials used *imputation*, in which some rule is used to fill in the missing data: last observation carried forward (8 trials), worst-case analysis (5 trials), and regression imputation including multiple imputation[5] (2 trials). Once imputation is done, the analysis proceeds along standard lines. Five further trials used modelling methods to construct a statistical model for the longitudinal data and thereby dealt with the missing data and the analysis in a single step. Finally, the methods used in two trials were unclear.

This paper compares the assumptions and practical implementation of the above methods, with the aim of establishing which are appropriate and in what circumstances. We also consider alternative models for regression imputation, and we describe new extensions to the modelling strategies that allow for a possible association between dropout and outcome, conditional on the observed data.

We illustrate the methods using data from a recent 12-month randomized controlled trial, comparing the effectiveness of two interventions for increasing physical activity.[6] As 40% of participants dropped out before the primary end-point at 12 months, it is important to assess the sensitivity of the conclusions to the analysis methods.

Data

The data come from a randomized controlled trial to assess the effectiveness of two primary-care-based interventions for promoting physical activity. In all, 1658 eligible participants age 45–64 who returned a postal baseline questionnaire were randomized to either brief negotiation (BN, n = 551), direct advice (DA, n = 544), or a control group (n = 563). Participants in the intervention groups were then invited by post to a health check with a health promotion specialist. At this health check, participants were asked to consent to the randomized trial, and if consent was given the appropriate intervention was immediately carried out. Comparisons between intervention and control groups must be performed using all randomized individuals. However, this paper focuses on the comparison of BN with DA. Because the allocated intervention was concealed from the health promotion specialist and patient until consent was given, this comparison may validly be restricted to the 585 participants who gave their consent and received 30 minutes of either BN (n = 302) or DA (n = 283).

The outcome variable of interest was physical activity calculated as kilocalories expended per kilogram bodyweight per week (kcals/kg/week). Physical activity was self-reported at baseline through a 4-week retrospective questionnaire and after 12 months from a 28-day logbook. Intermediate assessments at 3, 6, and 9 months were also carried out for the intervention groups only, using self-completed 7-day logbooks. As in the original analysis,[6] any participants who returned the baseline questionnaire but left blank answers to all questions regarding physical activity of an intensity $\geqslant$3 MET (work metabolic rate/resting metabolic rate) were assigned a baseline physical activity value of zero (34 [11.3%] of BN, and 21 [7.4%] for DA participants, $P = 0.11$). The baseline characteristics of interest to the investigation, were age, gender, education, employment status, home ownership, smoking, general health, physical health, and an indicator for zero baseline energy expenditure.

Missing data indicators are used in baseline covariates with missing values. Although this approach is known to be biased in observational studies,[7] it is appropriate in randomized trials because randomized treatment is independent of covariates.[3] Alternatively, missing baseline values could have been imputed either deterministically or by means of chained multiple imputations. See references 3 and 8 for more discussion.

In total, 231 (39%) participants receiving intervention failed to return the 12-month logbook (Figure 1). There was a mixture of intermittent missing data, that is when a participant fails to return a questionnaire but subsequent questionnaires are returned, and dropout, when once a participant misses a measurement time, no further questionnaires are returned.

The effect of treatment was assessed by the change in energy expenditure from baseline to the 12-month assessment. The distribution of energy expenditure was skewed and was therefore transformed by taking the logarithm after adding one kcal/kg/week (to avoid problems with baseline zeros). The trial analysis imputed missing 12-month outcomes using an imputation model constructed in the control group and found that the BN group reported 10% more activity change (about 3 kcal/kg/week) than the DA group, but the difference was not statistically significant. Further details about the trial methods and results can be found elsewhere.[6]

Methods for dealing with dropouts

Before describing the complete-case, imputation, and modelling methods in detail, we review a commonly used hierarchy of assumptions about missing data.

Assumptions

Many methods for dealing with dropouts make assumptions about how the probability of dropout relates to baseline covariates and

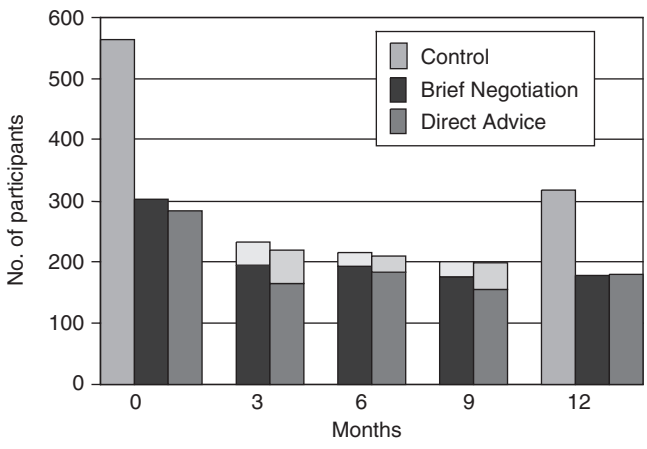

Figure 1 The number of participants remaining in the trial, the upper bar regions represent the number of participants with intermittent missing data at that particular measurement time

outcomes. Little and Rubin define the following situations.[9,10] Data are 'missing completely at random' (MCAR) if dropout does not depend on the baseline covariates or outcomes. More general is 'covariate-dependent missing completely at random' (CD-MCAR), where dropout can depend on baseline covariates but not on any outcomes. 'Missing at random' (MAR) further allows dropout to depend on any observed data, including observed post-baseline outcomes. For example, dropout might be determined by the observed outcome of the patient as the trial progresses.[11] Under any of the above conditions, dropout is often referred to as 'ignorable', since a likelihood-based analysis using only the observed data can be valid given the model is correctly specified. This technical definition of ignorable should not be misunderstood to mean one should ignore individuals with missing data (e.g. complete-case analysis). Similar definitions can be made for intermittent missingness.

'Missing not at random' (MNAR) is used to describe situations when the probability of dropout additionally depends on the unobserved outcomes, as well as on the observed data. This arises, for example, when individuals leave a study because of a sudden unforeseen deterioration or improvement in their condition. Parameter estimation from the observed data alone is typically biased, to an extent that depends on the strength of the relationship between the unobserved outcome and dropout. Missing not at random is also known as informative or non-ignorable dropout.

Complete-case analysis (CC)

A recent survey[4] of published randomized controlled trials revealed that the most common way to deal with dropouts is to omit them altogether and analyse only data from individuals who complete the trial. This CC analysis is valid if the data are missing completely at random, so that conclusions about the population of completers also apply to the population of patients who drop out. A CC analysis that is adjusted for baseline covariates is valid under the less-restrictive assumption of covariate-dependent missing completely at random. However, any CC analysis is potentially inefficient.

In the physical activity trial, we perform a CC analysis for the change in outcome from baseline to 12 months, using standard linear regression to adjust for baseline covariates.

Ad hoc imputation

A crude, yet popular imputation technique is last observation carried forward (LOCF), which makes use of baseline measurements and any observed intermediate measurements by carrying forward the last observation to the final time point for patients who drop out. Baseline carried forward (BCF) imputes missing outcomes with the baseline value even if intermediate measurements are available. These techniques assume that an individual's missing value follows the same distribution as that for previously measured or baseline values for that individual. The methods are sometimes claimed to be conservative[12] in non-progressive conditions,[13] but this is not necessarily the case when there are more missing data in one group or where selection bias may occur and operate differently between the two groups. An alternative approach is to impute with the worst-case value (WC), if such a value exists. The assumption is that dropout implies poor condition, and such imputations are often extreme, so this approach rarely leads to unbiased estimates of the intervention effects.[14]

In the physical activity trial, we impute the missing 12-month outcomes using each of the above techniques in turn, and then include all randomized individuals in a linear regression analysis for the change in outcome from baseline to 12 months, adjusting for baseline covariates as previously mentioned. For the WC method we impute zero for missing energy expenditure values at 12 months. Because 12-month outcomes are much larger than baseline outcomes, LOCF is even less appealing than usual, and this motivated the use of regression imputation in the actual published analysis.

Model-based imputation

In these analyses, an *imputation model* is constructed relating the outcome to covariates and possibly to intermediate values of the outcome. The imputation model is usually fitted to the individuals with the observed outcome, and used to predict the unobserved outcomes. This technique gives an unbiased estimate of the intervention effect under the CD-MCAR or MAR assumption, provided that the imputation model is correctly specified. However, use of a single imputation,[10] whether deterministic or stochastic, generally underestimates the standard errors.[15] *Multiple imputation* uses several stochastic imputations which are pooled together to provide a single estimate. Multiple imputation gives valid standard errors (under MAR) that incorporate the uncertainty about the imputed values: for further details, see reference 5.

For the analysis of the physical activity trial, we constructed two separate regression imputation models using baseline data. The first was constructed amongst completers in the intervention groups, regressing on all baseline covariates including intervention group. These imputations are group specific. We let RI and MI denote a single regression imputation analysis and a multiple imputation analysis, respectively. This model makes the CD-MCAR assumption that the unobserved outcomes in each group do not differ systematically from the observed outcomes in the same treatment group, once differences in baseline characteristics are taken into account.

It would further be possible to extend the RI and MI approaches to incorporate intermediate outcomes, which would make a more general assumption of MAR. An alternative approach is longitudinal modelling, described in the next section.

The investigators in the physical activity trial believed that individuals with missing outcomes were unlikely to have benefited from the intervention. This motivated a second regression imputation model which was constructed amongst completers in the *control* group, regressing on all baseline covariates except intervention group. This method was used in the primary analysis of the physical activity trial.[6] We let RI-c and MI-c denote these single and multiple imputation analyses. These approaches assume that the unobserved outcomes in each intervention group are comparable with the observed outcomes in the control group, again allowing for differences in baseline characteristics. This assumption does not fit neatly within the MCAR/MAR/NMAR framework and typically is a special case of NMAR. The implicit assumptions of this method may be implausible in many trials, especially when there are missing outcomes in the control

group. As a sensitivity analysis, we also used a pooled intervention model, (as RI/MI but with no intervention effect in the imputation model), and this would also be applicable in a trial without a control group.

For practical analyses five imputed data sets are considered enough for multiple imputation. For comparing estimated intervention effects between methods, we needed to minimize Monte Carlo error. Because of the large proportion of missing data, we found that 1000 imputed data sets were required to give comparable results to the bivariate model described below.

Longitudinal model with random effects (RE)

A longitudinal model with random effects[16] allows all observed repeated measures to be included in the analysis and is valid under the MAR assumption given the model is correctly specified. In the past, longitudinal data models have required balanced designs, in which all individuals are observed at a set of pre-specified time points. A longitudinal random effects model is less restrictive and can be used for unbalanced designs. See Appendix A for our model definition.

Baseline data can be used in three different ways. Firstly, the outcome can be defined as change from baseline. Secondly, the baseline can be used as a covariate. Thirdly, the baseline value can be included as the first observation of the repeated measures. We chose the third option because it is able to include all participants in the longitudinal data analysis, including participants with only baseline data. This approach accords better with the intention-to-treat principle, although in practice the inclusion or exclusion of individuals with only baseline data affects only the within-group means and not the treatment effect.[17] Unlike the previously described approaches, a normality assumption is made for baseline outcomes.

We assume a saturated model for the mean outcomes with distinct parameters for each intervention and measurement time combination, except that the intervention is not allowed to affect baseline outcome. All baseline covariates, including dummy variables, were centred by subtracting their mean, so that the treatment-time parameters are interpretable as adjusted means. We performed checks for whether the effect of baseline covariates and response variance varied across the measurement times, and found that they were different at baseline, intermediate measurement times (3, 6, 9 months), and 12 months, which is consistent with the different assessment procedures. Hence, we allow for distinct parameters for each baseline covariate and assessment procedure combination, and the response variances to be different at 0, and 3, 6, 9 months, and 12 months. Adjustment for the indicator of baseline zero energy expenditure is made at post-baseline outcomes. Change in efficacy assessments at the final time point can be extracted from the model parameters. We consider two models; *RE*, which includes all intermediate values as described above, and *RE-biv*, which is a bivariate model for baseline and final outcome only, thus excluding the intermediate values. For such data it may be preferable to use an unstructured covariance matrix, but in fact results were very similar to results from the random effects model. We present the results from a random effects model because it generalises more easily in the joint models discussed below.

Joint models for longitudinal data and non-ignorable dropout

When MAR is not assumed, so that the data are MNAR, a joint model of the outcomes and dropout information must be considered. This can be done by combining a model for the outcomes, such as the longitudinal model described above, with a 'selection' model for dropout conditional on outcome.

We consider two flexible approaches for modelling the relationship between the outcomes and dropout, classified as '*random effect-dependent dropout*' in Little's[9] framework. The underlying assumption is that dropout depends on some linear combination of the random effects from the longitudinal model. This can allow, for example, individuals with lower (or higher) than average outcomes to be have a greater risk of dropout. See Appendix B for model descriptions.

Our first approach models whether individuals drop out before 12 months via logistic regression allowing for dependence on baseline covariates and on the value of the individual random effects at the final measurement time from the longitudinal model; see equation (A2) in Appendix B. The model can be adapted to allow for other choices for the relationship between dropout and outcomes, for example, it may be likely that dropout also depends on the rate of decline in the outcomes. One major potential source of bias arises when missingness is more informative in one group than the other and we may wish to allow the relationship between dropout and outcome to be different between treatment groups.

A second approach uses actual times of dropout and models dropout at each measurement time either using a logistic or time-to-event model. We use a proportional hazards model which allows dropout at time t to depend on baseline covariates and the individual random effects from the longitudinal model at time t; see equation (A3) in appendix B. Like (A2), this model is flexible and various associations between dropout and the random effects can be incorporated.

In application to the physical activity trial we use the same longitudinal model as described in the above subsection to model the outcomes and combine with each of the conditional dropout models described by (A2) and (A3) to formulate two joint models. We let JM-log represent the logistic joint model and let JM-ph represent proportional hazards joint model. To select covariates for the conditional dropout models, we performed simple stepwise procedures in models (A2) and (A3) with the association parameter, γ, set to zero.

Parameter estimation for the joint models is not available in standard statistical software and is fairly computationally intensive. We used WinBUGS[18,19] to estimate parameters in JM-log, see program in Appendix C, and an EM algorithm[20–22] to estimate parameters in JM-ph. Nested models within each framework are compared via their log-likelihoods.

Results

In all 583 participants were randomized to BN or DA. Figure 1 shows the number of patients observed at each measurement time, and the number of participants not yet dropped out at each time. There is no significant difference in dropout over time between intervention groups ($P = 0.35$).

There are no significant differences in baseline covariates between the two intervention groups. However, age, the

percentage of homeowners, the percentage of individuals who are permanently sick or disabled and the baseline physical activity levels differ between completers and dropouts (Table 1).

We investigated predictors of dropout before the final 12-month assessment and the predictors of dropout over the course of the study using a multivariate logistic model and a Cox proportional hazards model respectively, shown in Table 2 (columns 1–4). Those who are permanently sick or disabled, those who are younger and those with low baseline physical activity levels are less likely to complete the 12 month logbook.

Complete-case analysis

Figure 2 shows the progression of the two intervention groups and the control group by their observed mean change in log energy expenditure as a function of time. Each average is over patients with an observed outcome at the measurement time. All groups have increasing physical activity levels over time, with a slower rate towards the end of the study. The overall increase in log energy expenditure at 12 months within each intervention group is 0.88 (95% CI: 0.73, 1.03) (BN) and 0.60 (95% CI: 0.47, 0.73) (DA), Table 3. These are calculated from 176 (BN) and 178 (DA) completers, Figure 1. Both unadjusted and adjusted CC analyses yield significant between-group differences (Table 3), indicating a greater increase in physical activity in the brief negotiation group than the direct advice group, amongst participants who completed the trial. However, this method violates the intention-to-treat principle by excluding roughly 40% of participants with missing outcomes.

The difference in the results between the unadjusted and adjusted analyses is mainly caused by the adjustment for baseline energy expenditure and the zero-baseline indicator in the adjusted analysis.

Ad hoc imputation

Table 3 also reports the results for the change in physical activity at 12 months using imputation techniques to deal with dropouts, as described previously. The methods allow all individuals to be included in the analysis. Due to the increasing trend in physical activity over the 12 months (Figure 2), imputing with the baseline or the last observed value decreases the mean changes in physical activity compared with the CC analysis (Table 3). However, the retrospectively reported baseline measurements clearly are not comparable with the logbook-based post-baseline measures (Figure 1), and imputing with these values seems inappropriate. Most dropouts occur between baseline and 3 months, so even LOCF uses a majority of baseline measures to impute the 12-month outcome. Imputing using the worst-case value is the only approach that yields negative mean changes in physical activity at 12 months in both intervention groups. These methods produce widely

Table 1 Comparisons in baseline characteristics between completers and dropouts by intervention group. Values are means (standard deviations) unless otherwise stated

	Brief negotiation			Direct advice		
Characteristic	Completers	Dropouts	*P*-value[a]	Completers	Dropouts	*P*-value[a]
n	**176**	**126**		**179**	**104**	
Age (years)	55.1 (5.7)	54.1 (5.1)	0.18	55.5 (5.8)	54.5 (5.8)	0.15
Men %	41.5	43.7	0.71	48.6	44.2	0.48
Education %						
Higher Qualification	18.8	10.3	0.13	10.6	8.7	0.62
A level or equivalent	5.1	4.0		5.0	1.9	
O level or equivalent	15.9	10.3		15.1	12.5	
Other	18.8	26.2		19.6	26.0	
None	38.1	43.7		45.3	47.1	
Employment status %						
Unemployed	28.4	24.6	0.46	26.3	30.8	0.42
Home ownership %						
Owned/mortgaged	86.9	79.4	0.08	88.8	82.7	0.14
Smokers %	15.9	23.0	0.25	23.5	29.8	0.03
General health %						
Good	72.7	71.4	0.21	76.5	74.0	0.57
Fair	26.7	24.6		20.7	21.2	
Bad	0.6	3.2		2.2	4.8	
Physical health %						
Permanently sick/disabled	3.4	10.3	<0.01	1.7	6.7	0.16
Log energy expenditure (kcal/kg/week)	2.4 (1.2)	2.1 (1.3)	0.01	2.6 (1.2)	2.2 (1.2)	0.11
Zero energy expenditure[b] %	6.8	17.5	<0.01	3.4	14.4	<0.01

[a] *P*-values are for the within group comparison between completers and dropouts.

[b] No recorded physical activities ⩾3 MET.

Table 2 Unadjusted and adjusted odds ratios for dropout before 12 months (logistic regression); unadjusted and adjusted hazard ratios for dropout over 12 months (proportional hazards); adjusted odds ratios for joint model JM-log (described in text) and adjusted hazard ratios for joint model JM-ph (described in text). Covariates were selected from stepwise regressions in the dropout models

	Dropout models				**Joint models for longitudinal data and non-ignorable dropout**	
	Logistic regression		Proportional hazards		JM-log	JM-ph
	Unadjusted odds ratio (95% CI)	Adjusted odds ratio (95% CI)	Unadjusted hazard ratio (95% CI)	Adjusted hazard ratio (95% CI)	Adjusted odds ratio (95% CI)	Adjusted hazard ratio (95% CI)
Baseline physical activity less than one	1.68	1.65	1.46	1.43	1.71	1.41
30-minute occasion per week	(1.20, 2.37)	(1.17, 2.34)	(1.13, 1.90)	(1.10, 1.86)	(1.15, 2.54)	(1.16, 1.72)
Permanently sick or disabled	3.66	3.67	2.06	2.14	3.68	2.14
	(1.64, 8.19)	(1.61, 8.33)	(1.29, 3.30)	(1.35, 3.40)	(1.56, 9.23)	(1.46, 3.13)
Age (1 year change)	0.97	0.96	0.98	0.97	0.95	0.97
	(0.94, 1.00)	(0.93, 0.99)	(0.95, 1.00)	(0.95, 1.00)	(0.92, 0.98)	(0.95, 0.98)
Random effects at 12 months					0.98	—
					(0.76, 2.18)	
Random effects at dropout time *t*					—	0.96
						(0.76, 1.22)

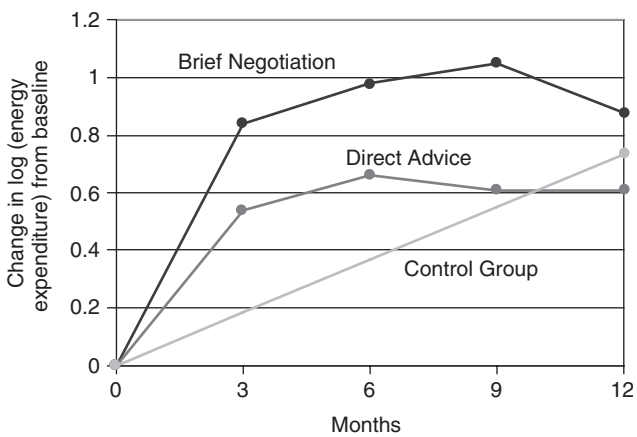

Figure 2 The observed mean change in physical activity among those patients with an observed response at the measurement time

differing results, although in each case the adjusted intervention effect lacks statistical significance. The standard errors are also likely to be too small because the imputed data are assumed to be true.

Model-based imputation

The mean changes in outcome from baseline to 12 months increase under imputation methods RI and MI, because dropouts tended to have baseline covariates more associated with larger changes (Table 3). A smaller increase in the mean change is observed under imputation schemes RI-c and MI-c. These results agree closely with the previous analysis of the trial,[6] although that analysis computed standard errors by adjusting the single-imputation standard error. The mean change in the DA group is fairly robust to the regression imputation strategies, indicating similarities between the observed DA group and the control group.

As expected, the multiple imputation strategies (MI, MI-c) yield similar point estimates but with increased confidence intervals compared with the single regression imputations (RI, RI-c respectively). The adjusted between-group differences from RI and MI are comparable with the adjusted complete-case analysis, as would be expected since both assume CD-MCAR. The between-group differences from RI-c and MI-c are smaller because the missing outcomes have been replaced using an imputation model with no intervention effect. Similar results were obtained by fitting an imputation model with no intervention effect to the intervention group data: this shows that the critical issue is whether the imputation model is allowed to differ between the treatment groups.

Longitudinal random effects model

Figure 3 shows the observed mean changes in energy expenditure over the study period, along with fitted profiles from the adjusted longitudinal data analysis. The latter represents the estimated mean changes in outcomes using all observed and unobserved measures, obtained using the estimated parameters from the longitudinal random effects model. Note that individuals who drop out immediately after baseline are excluded from the calculation for the observed profiles, which has caused a large difference between the observed and fitted profiles in the DA group.

Table 3 includes the results from the longitudinal data analysis, RE, and the bivariate-data analysis, RE-biv. We considered both in order to assess the impact of including the intermediate measurements. In fact, there is little difference between the results indicating that in this trial, conditional on baseline information, intermediate measurements have little effect on the 12 month outcome.

The RE group mean changes (Table 3) are adjusted for both baseline covariates and intermediate values and are larger than those from the CC analysis mainly because the dropouts have baseline covariates which tend to be associated with larger changes. The group mean changes from RE also differ from

Table 3 Mean changes in log-transformed physical activity after 12-months by intervention group, under various statistical analyses

	Mean change at 12 months (95% CI)		Difference (Brief negotiation—Direct advice) (95% CI)	
Statistical method	Brief Negotiation N = 302 (176 observed)	Direct advice N = 285 (178 observed)	Unadjusted	Fully adjusted[a]
Complete cases (CC)	0.88 (0.73, 1.03)	0.60 (0.47, 0.73)	0.28 (0.04, 0.53)	0.18 (0.01, 0.35)
Imputation methods				
Last observation carried forward (LOCF)	0.71 (0.58, 0.84)	0.51 (0.39, 0.63)	0.13 (−0.03, 0.29)	0.13 (−0.02, 0.28)
Baseline carried forward (BCF)	0.51 (0.40, 0.63)	0.39 (0.28, 0.49)	0.20 (0.02, 0.38)	0.08 (−0.06, 0.22)
Worst case (WC)	−0.35 (−0.57, −0.13)	−0.44 (−0.65, −0.24)	0.09 (−0.21, 0.40)	0.00 (−0.27, 0.27)
Regression imputation (RI)	1.01 (0.86, 1.15)	0.71 (0.63, 0.79)	0.30 (0.12, 0.48)	0.17 (0.07, 0.26)
Regression imputation (RI-c)	0.91 (0.78, 1.04)	0.69 (0.56, 0.81)	0.23 (0.05, 0.41)	0.10 (− 0.00, 0.21)
Multiple imputation (MI)	1.01 (0.86, 1.16)	0.71 (0.57, 0.85)	0.30 (0.08, 0.53)	0.17 (0.00, 0.35)
Multiple imputation (MI-c)	0.91 (0.77, 1.05)	0.68 (0.55, 0.82)	0.23 (0.01, 0.44)	0.10 (− 0.05, 0.26)
Longitudinal modelling methods	**(Covariate adjusted)**		**(Baseline-outcome adjusted)**	**(Fully adjusted[a])**
Bivariate random effects model (RE-biv)	0.95 (0.81, 1.09)	0.78 (0.64, 0.91)	0.17 (0.00, 0.35)	0.17 (0.00, 0.34)
Longitudinal random effects model (RE)	0.96 (0.82, 1.10)	0.78 (0.64, 0.93)	0.16 (0.00, 0.32)	0.18 (0.01, 0.35)
Joint model JM-log[b]	0.97 (0.82, 1.12)	0.79 (0.63, 0.94)	0.17 (0.00, 0.34)	0.18 (0.01, 0.35)
Joint model JM-ph[b]	0.98 (0.84, 1.11)	0.80 (0.65, 0.94)	0.17 (−0.01, 0.34)	0.18 (0.02, 0.34)
Joint model JM-log with group-interaction[c]	0.96 (0.80, 1.12)	0.80 (0.63, 0.96)	0.16 (−0.04, 0.37)	0.17 (−0.04, 0.37)

[a] Adjusted for baseline energy expenditure, age, gender, health status, employment, education, and home ownership.

Joint model JM-log: logistic model relates dropout before 12 months to random effects at 12 months. Fitted in WinBUGS.

Joint Model JM-ph: proportional hazards model relates dropout at time *t* to random effects at time *t*. Fitted by Maximum Likelihood.

Joint Model JM-log with group-interaction: as JM-log above with an interaction term between the random effects and intervention group.

[b] Dropout model adjusts for covariates age, permanent sickness or disability, and low baseline activity.

[c] Dropout model adjusts for covariates age, permanent sickness or disability, low baseline activity, and intervention group.

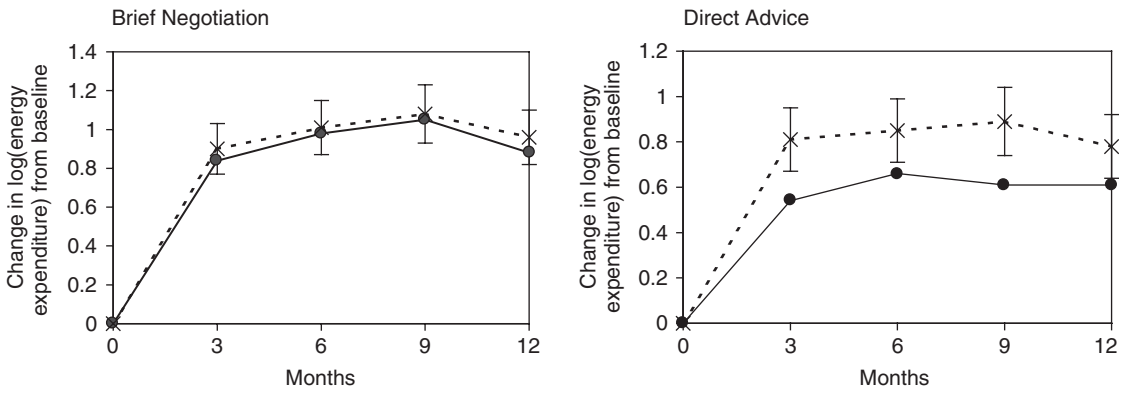

Figure 3 The observed mean change in log energy expenditure (solid lines); and the estimated mean changes from the longitudinal model with random effects (broken lines) for the two intervention groups

those produced by the RI and MI analyses. These approaches assume MAR, but despite this, the group mean changes are not expected to be identical because the underlying models are different. Specifically, the RI and MI approaches use the group-specific baseline and final outcome measures to obtain the group mean change; whereas, the RE method uses the overall mean baseline outcome and the group-specific 12-month outcome, because we did not allow the baseline outcome to vary between the two groups in the RE model. However, as we would expect, the adjusted between group differences from these MAR analyses are almost identical (Table 3).

Joint models for longitudinal data and non-ignorable dropout

The results for the mean change in physical activity at 12 months from the two joint models, JM-log and JM-ph, are shown in Table 3. The results are very similar to those observed from the RE analysis alone. The RE model is the same as the joint models with $\gamma = 0$, and so we could expect it to give smaller standard errors than the joint model analysis. However, we did not observe this in our results. In both joint models, there remains a highly significant effect of low baseline physical activity on dropout (Table 2). However, there are no significant associations between dropout and the random effects from the longitudinal model, suggesting that dropout may be independent of the unobserved outcomes, under the assumptions of these models.

To explore the importance of correct model choice, we excluded baseline energy expenditure from the conditional dropout model. The association between dropout and the random effects became significant and negative, indicating that participants with lower than average energy expenditure are more likely to drop out (JM-ph: HR = 0.82, 95% CI: 0.71, 0.95). However, the adjusted between group difference was unaffected (adjusted between-group difference = 0.18, 95% CI: 0.00, 0.36). Other models for association were considered, including an interaction term between the random effects and intervention in the conditional dropout model (Table 3), and allowing dropout to depend on the rate of change in the physical activity responses. These did not improve the likelihood or affect the estimated intervention effect.

Discussion

Three principles should be kept in mind when analysing a longitudinal study with missing outcomes. First, each individual should be included in the analysis if at all possible, especially in a randomized trial where this is required by the intention-to-treat principle. Secondly, the assumptions behind the analysis should be stated and should be assessed both from subject-matter knowledge and, where possible, from the data. Thirdly, suitable sensitivity analyses should be performed to assess the impact of departures from the assumptions.

We have considered three families of techniques for dealing with missing data. CC analysis is the most commonly used technique, but it does not respect the intention-to-treat principle. It cannot be recommended when an outcome is repeatedly measured, since it discards data on observed intermediate outcomes from individuals with missing final outcome. However, for a singly measured outcome, model-based imputation RI/MI and bivariate modelling RE-biv give essentially the same treatment effects as CC analysis. In this situation, the fact that model-based imputation and bivariate modelling analyses conform to the intention-to-treat principle is only a cosmetic advantage.

Ad hoc imputation techniques vary widely in their assumptions; in our data, they also varied widely in their estimates of the mean change over time and of the intervention effect. In practice we suspect that techniques such as last observation carried forward (LOCF) are adopted with little thought as to their appropriateness. Our results therefore suggest that serious errors may be incurred by using such crude imputations.

In particular, LOCF has been advocated as a conservative approach to trials, such as this one, in which mean outcomes improve over time. It is conservative for the estimated change over time, but it is not necessarily conservative for the difference between the groups. In our example, mean physical activity at baseline was substantially lower than mean physical activity at follow-up. Change over time is therefore under-estimated whenever baseline values are carried forward, and the treatment comparison will be biased against whichever group has more baseline values carried forward. This may not be the

group with more missing data overall, so the direction of the bias is not easily predictable.

The longitudinal modelling techniques are closely related to one another, since they are all based on a model relating the repeated measures to an underlying individual trend over time. The repeated measures model assumes that data are missing at random—that is, that dropout may depend on previously observed outcomes, but not additionally on unobserved outcomes. The joint model instead assumes that dropout may depend on the underlying individual trend. In many situations the latter assumption is at least as plausible as the missing at random assumption, and the joint model therefore serves as a valuable sensitivity analysis. In the physical activity trial, the assumptions of the joint model were plausible and the results were very similar to the repeated measures model, thus supporting the results of the repeated measures model.

Other joint modelling techniques are possible. Verzilli and Carpenter[22] modelled the association between dropout and the repeated measures through the observed and unobserved values of the repeated measures themselves, rather than through the random effects. Such models can be used for sensitivity analysis by allowing the association parameter to take on a range of values. By contrast, the joint models presented directly estimates the association between dropout and the repeated measures through the random effects. A significant association suggests evidence of non-ignorable dropout, provided the model assumptions are correct.

Analyses assuming missing at random can be found both in the imputation family (regression imputation allowing for randomised group), and in the modelling family, (the longitudinal random effects model). It is not surprising that these methods gave very similar estimates of the adjusted between-group difference in our example. A fourth class of methods, weighting by the inverse probability of missingness, can also give valid results under CD-MCAR, see for example reference Dowrick,[23] and can be extended to be valid under MAR,[24] but are inefficient without a complex optimisation procedure.[25]

The longitudinal modelling and imputation families did differ in one important respect: imputation based on the control group gave substantially different results from imputation allowing for randomised group and from the modelling approaches. Imputation based on the control group assumes that individuals with missing outcomes did not derive benefit from the intervention, while the latter assumes that they did. It is not possible to use the data to assess which assumption is correct, since the assumptions relate entirely to unobserved data. The choice must therefore be based on subject matter considerations. The choice also depends on the aim of analysis—imputation ignoring randomized group would be inappropriate if the aim was to estimate efficacy rather than effectiveness.[26]

In the physical activity trial, we found that low baseline physical activity, age, and permanent sickness or disability are predictors for dropout, but it is unlikely that dropout depends on unobserved physical activity responses. However, the joint models have allowed scope to explore non-ignorable missing data mechanisms, leading to more confident effect estimates. Our analysis shows that the conclusions from the trial depend critically on beliefs about the dropouts. If one believes that dropouts were similar to other members of the same intervention group then brief negotiation to increase physical activity is significantly more effective than the direct advice intervention (adjusted between group difference = 0.18, 95% CI: 0.02, 0.35). However, if we believe dropouts have similar activity levels to individuals in the control group, the effect falls below significance (adjusted between group difference = 0.10, 95% CI: −0.05, 0.26) but still suggests a benefit.

In conclusion, both the imputation family and the longitudinal modelling family offer acceptable ways to analyse trials with data missing at random. A good analysis should consider more extreme possibilities, and we have demonstrated ways to do this: in the imputation family, by altering assumptions about the impact of intervention; in the longitudinal modelling family, by altering assumptions about the missingness mechanism. The choice between these methods can only arise from an understanding of the context.

KEY MESSAGES

- A range of statistical methods is available to deal with missing outcome data; all methods have underlying assumptions, which vary widely in their plausibility.
- *Ad hoc* imputation methods should be avoided in favour of methods with more plausible assumptions, such as multiple imputation and longitudinal modelling.
- Suitable sensitivity analyses should be performed to assess the impact of departures from the assumptions.

References

[1] Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999;**319:**670–74.

[2] Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J Clin Epidemiol* 2003;**56:**28–37.

[3] White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomised trials. *Stat Med* (in press).

[4] Wood A, White I, Thompson S. Are missing outcome data adequately handled? A review of published randomised controlled trials. *Clin Trials* (in press).

[5] Schafer JL. Multiple imputation: a primer. *Stat Meth Med Res* 1996;**8:**3–15.

[6] Hillsdon M, Thorogood M, White I, Foster S. Advising people to take more exercise is ineffective: a randomized controlled trial of physical activity promotion in primary care. *Int J Epidemiol* 2002;**31:**808–81.

[7] Greenland S, Finkle W. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 1995;**142:**1255–64.

[8] Buuren S van, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999;**18:**681–94.

[9] Little R. Modeling the dropout mechanism in longitudinal studies. *J Am Stat Assoc* 1995;**90:**1112–21.

[10] Little R, Rubin D. *Statistical Analysis with Missing Data*. New York: Wiley, 1987.

[11] Murray GD, Findlay J. Correcting for bias caused by drop-outs in hypertension trials. *Stat Med* 1988;**7:**941–46.

[12] Shih W. Problems in dealing with missing data and informative censoring in clinical trials. *Curr Control Trials Cardiovasc Med* 2002;**3:**4.

[13] Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Stat Med* 2001;**20:**3931–46.

[14] Unnebrink K, Windeler J. Sensitivity analysis by worst and best case assessment: is it really sensitive? *Drug Inf J* 1999;**33:**835–39.

[15] Myers W. Handling missing data in clinical trials: an overview. *Drug Inform J* 2000;**34:**525–33.

[16] Laird N, Ware J. Random-effects models for longitudinal data. *Biometrics* 1982;**38:**963–74.

[17] White IR, Moodie E, Thompson SG, Croudace T. A modelling strategy for the analysis of clinical trials with partly missing longitudinal data. *Int J Meth Psych Res* 2003;**12:**139–50.

[18] Spiegelhalter DJ, Thomas A, Best NG, Lunn D. *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge, available from www.mrc-bsu.cam.ac.uk/bugs 2002.

[19] Carpenter J, Pocock S, Lamm C. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. *Stat Med* 2002;**21:**1043–66.

[20] Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000;**1:**465–80.

[21] Dobson A, Henderson R. Diagnostics for joint longitudinal and dropout time modelling. *Biometrics* 2003;**59:**741–51.

[22] Verzilli CJ, Carpenter JR. A Monte-Carlo EM algorithm for random-coefficient-based dropout models. *J Appl Stat* 2002;**29:**1011–21.

[23] Dowrick C, Dunn G, Ayuso-Mateos JL *et al*. Problem solving treatment and group psychoeducation for depression: multicentre randomised controlled trial. *BMJ* 2000;**321:**1450–54.

[24] Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995;**90:**106–21.

[25] Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc* 1995;**90:**122–29.

[26] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis* 1967;**20:**637–48.

[27] SAS Institute Inc. *SAS/STAT User's Guide, Version 8*. Cary, NC: SAS Institute Inc, 2000.

[28] Rasbash J, Browne W, Goldstein H *et al. A User's Guide to MLwiN (Version 2.1d)*. Multilevel Models Project, Institute of Education University of London, 2002.

Appendix A

Longitudinal model with random effects

Let m be the total number of individuals, with n_i measurements made on individual i for $i = 1, \ldots, m$. The set of outcomes will be denoted by $\{y_{ij}: j = 1,\ldots,n_i,\ i = 1,\ldots,m\}$ recorded at the set of times $\{t_{ij}: j = 1,\ldots,n_i,\ i = 1,\ldots,m\}$. Throughout we assume independence between individuals.

We will assume that each observation comes from a Normal distribution, and measurements made on individuals within groups and at the same time points are from the same distribution. The general model for an individual's outcome will be of the form:

$$Y_{ij} = \mu_i(t_{ij}) + U_{1i} + U_{2i}\, t_{ij} + Z_{ij}. \qquad \text{(A1)}$$

The mean outcome $\mu_i(t_{ij})$ can be described by a linear model $\mu_i(t) = x_{1i}(t)\ '\beta_1$ where $x_{1i}(t)$ is a p_1-element vector of known individual baseline covariates, possibly-time varying, and β_1 is a vector of dimension p_1 of unknown regression coefficients. The U_{1i} and $U_{2i}\, t_{ij}$ are random intercept and slope terms that allow for variation of the mean outcome profile and the dropout rate respectively. We assume (U_{1i}, U_{2i}) come from a bivariate zero-mean Normal distribution with variances σ_1^2 and σ_2^2 respectively, and correlation ρ. Finally, $Z_{ij} \sim N(0, \sigma_{ej}^2)$ are mutually independent and represent measurement error. Estimation of this type of model requires the use of an optimisation algorithm and is available in standard software.[27,28] We used an EM algorithm.

Appendix B

Joint model for longitudinal data and non-ignorable dropout

A joint model for the longitudinal data and non-ignorable dropout is formulated by combining the random effects longitudinal model in (A1) with a model relating dropout to the outcomes. We assume that dropout depends on a linear combination of the random effects from the longitudinal model. Model JM-log uses a logistic regression model for dropout before 12 months:

$$\begin{aligned}&\text{logit}(P[i \text{ drops out before final time } t \mid U_1, U_2])\\ &\quad = x_{2i}'\beta_2 + \gamma\,(U_{1i} + U_{2i}\, t)\end{aligned} \qquad \text{(A2)}$$

where x_{2i} is a p_2-element column vector of baseline covariates, β_2 is a p_2-element parameter vector, and γ is the association parameter which links dropout with a function of the outcomes. A further term δU_{2i} can be added to allow dropout to depend on the rate of decline in the outcomes, or adding the term $\delta(U_{1i} + U_{2i}t)\, R_i$ allows for the association between dropout and outcome to be different between treatment group, where R_i represents the group.

Model JM-ph uses a proportional hazards model for dropout at each measurement time:

$$h(t) = h_0(t)\, \exp(x_{2i}'\beta_2 + \gamma\,(U_{1i} + U_{2i}\, t)), \qquad \text{(A3)}$$

where the function $h_0(t)$ is a baseline hazard which applies for all patients. Note that in the physical activity trial the dropout times are discrete, so we assume a constant baseline hazard between dropout times. Again, additional terms may be added as appropriate.

Appendix C

Example of WinBugs model code

The example model shown below excludes baseline covariates and also allows the response variance to be different for each assessment procedure, that is, at baseline, intermediate follow-up

and the final measurement time. The code was used to produce the JM-log results in Table 3 in the baseline-outcome adjusted column.

```
model {
#M = number of participants.
#n = max number of repeated measures including
baseline outcome.
#y: matrix (M*n) of repeated measures.
#miss: indicator 0/1 for dropout before final
measurement time.
#group: indicator 0/1 for intervention groups.
for (i in 1:M){
#baseline outcome

  y[i,1] ~ dnorm(mu[i,1], tau[1])

  mu[i,1] <-beta[1] + u[i,1] + u[i,2]*t[1]

#intermediate outcomes

  for(j in 2:(n-1) ){

  y[i,j] ~ dnorm(mu[i,j], tau[2])

mu[i,j] <-beta[j] + beta[n−1 + j]*group[i]
         + u[i,1] + u[i,2]* t[j]}

#final outcome
  y[i,n] ~ dnorm(mu[i,1], tau[3])

  mu[i,n] <-beta[n] + beta[n−1 + n]*group[i]
           + u[i,1] + u[i,2]* t[n]}
#random effects

  u[i,1:2] ~ dmnorm(mu.u[1:2], sigma[1:2,1:2])

#dropout process

  miss[i] ~ dbern(p[i])

  logit(p[i])<- g[1] + g[2]*(u[i,1] + u[i,2] *t[n])

  }

  #priors

  for(k in 1:(n + n−1)) {beta[k] ~ dnorm(0, 0.001)}

  for(j in 1:3) {tau[j] ~ dgamma(0.05, 0.001)}

  for(j in 1:2) {g[j] ~ dnorm(0, 0.0001)}

  sigma[1:2,1:2] ~dwish(R[1:2,1:2], 2)

  }
```