A modelling strategy for the analysis of clinical trials with partly missing longitudinal data

IAN R. WHITE¹, ERICA MOODIE², SIMON G. THOMPSON¹, TIM CROUDACE³

¹MRC Biostatistics Unit, Cambridge, UK

²University of Seattle, USA

³University of Cambridge, UK

ABSTRACT *Standard statistical analyses of randomized controlled trials with partially missing outcome data often exclude valuable information from individuals with incomplete follow-up. This may lead to biased estimates of the intervention effect and loss of precision. We consider a randomized trial with a repeatedly measured outcome, in which the value of the outcome on the final occasion is of primary interest. We propose a modelling strategy in which the model is successively extended to include baseline values of the outcome, then intermediate values of the outcome, and finally values of other outcome variables. Likelihood-based estimation of random effects models is used, allowing the incorporation of data from individuals with some missing outcomes. Each estimated intervention effect is free of non-response bias under a different missing-at-random assumption. These assumptions become more plausible as the more complex models are fitted, so we propose using the trend in estimated intervention effects to assess the nature of any non-response bias. The methods are applied to data from a trial comparing intensive case management with standard case management for severely psychotic patients. All models give similar estimates of the intervention effect and we conclude that non-response bias is likely to be small.*

Key words: random effects modelling, multilevel modelling, incomplete data, missing values, MAR, longitudinal data, clinical trials

Introduction

Even with the best study design and data collection procedures, some outcome data are missing in nearly every clinical trial and longitudinal study. Missing data are an important potential source of bias because subjects with missing data may differ systematically from those with complete data. Since the 1970s, the volume of methodological literature on analysis of incomplete data has increased greatly (Little and Rubin, 1987; Schafer, 1997; Molenberghs et al., 1999) due in part to advances in computer software.

In simple data sets with very small amounts of missing data, it may be possible to derive meaningful conclusions from worst-case scenarios. For example, in a randomized controlled trial with a binary outcome, all missing data in the experimental arm may be replaced with treatment failures whilst all missing data in the control arm are replaced with treatment successes (Molenberghs, Kenward and Goetghebuer, 2001; Shadish et al., 1998). Commonly, however,

meaningful conclusions require making assumptions about the nature of the missing data. Under the 'missing-at-random' (MAR) assumption, the probability of an observation being missing may depend on observed variables (including previous values of the same variable) but may not depend on the true (unobserved) value of the variable conditional on the observed variables (Little and Rubin, 1987). For example, if patients whose symptoms are found to have worsened at a routine follow-up are removed from further follow-up, then MAR would hold provided that symptom severity is recorded at routine follow-ups (Murray and Findlay, 1988). Many common methods of analysis, however, use only individuals with complete data: this implicitly makes the more restrictive 'missing-completely-at-random' (MCAR) assumption, that the probability of an observation being missing does not depend on any observed or unobserved variables (Little and Rubin, 1987). If neither MCAR nor MAR holds – for example, if the

probability that a patient is not interviewed depends on his or her current disease severity – then the data are called ‘informatively missing’ (IM).

This paper focuses on randomized trials, although our methods could be extended to observational studies. We consider a trial whose outcome is measured repeatedly over time but is subject to missing values when attempts to interview patients failed. A suitable class of models for partly missing repeated measures data is the random effects models (Laird and Ware, 1982), also known as linear mixed models (Verbeke and Molenberghs, 1997), multilevel models (Goldstein, 1995) or random regression models (Gibbons et al., 1993). These models apply generally for data with an underlying hierarchical or clustered structure. A particular strength of random effects modelling is that it yields statistically efficient parameter estimates by using all observations in a data set, even when some subjects provide incomplete data or when observation times differ between individuals (Anderson, 1957; Everitt, 1998). Random effects modelling does not assume observations to be independent, thus allowing correlation of individuals’ measurements over time. Using restricted maximum likelihood (REML) estimating procedures (Cook, 1998), estimates of both fixed effects and variances are free of non-response bias when data are MAR.

An unexplored aspect of random effects models is the relationship between the choice of model and the MAR assumption. As well as the final value of the outcome of interest, it is possible to model its baseline values, its intermediate values, and concurrent values of other outcomes. These alternative statistical models have different MAR assumptions and therefore may suffer differing degrees of non-response bias.

Aims of this paper

This paper defines a sequence of analyses of a longitudinal randomized trial in which the value of the outcome on the final occasion is of primary interest. We start with simple two-group comparisons and successively add more information into the model until a comprehensive random effects model is reached. This sequence of analyses is used to offer increasing protection against the possibility of non-response bias as well as to quantify the treatment effect with the greatest precision.

We illustrate these methods using data from the UK700 trial, which compared intensive case

management with standard case management for patients with severe psychotic illness (Burns et al., 1999). Data were available from interviews at baseline, one year and two years. The follow-up rate was 86%, but some outcome variables had nearly a third of values missing due to incomplete interviews. This raised concerns over non-response bias. The aim here is to examine differences in estimates of the intervention effect and their precision, comparing traditional statistical analyses (t-tests and analysis of covariance) with a sequence of random effects models.

The UK700 trial

The UK700 trial was conducted in four inner-city areas. Patients aged 18–65 years were recruited if they had been diagnosed with a psychotic illness and had at least two psychiatric hospital admissions, the most recent within the previous two years. Patients were randomly allocated to either standard case management (caseload of 30–35 patients) or intensive case management (caseload of 10–15 patients). Details of patient selection and randomization have been published elsewhere (Burns et al., 1999).

The pre-specified primary outcome was the number of days in hospital for psychiatric reasons over the two years of the study; this was also recorded for the two years prior to baseline. Five pre-specified secondary outcomes were also recorded during interviews at baseline, year 1, and year 2: psychopathology, social disability, quality of life, unmet needs, and patient satisfaction with health services. For illustrative purposes only, two of these secondary outcomes will be considered here: psychopathology, which had the lowest proportion of missing values, and satisfaction with care, which had the highest proportion. Psychopathology was assessed with the Comprehensive Psychiatric Rating Scale (CPRS) (Asberg et al., 1978), which records 65 symptoms on a four-point rating scale (0–3). The responses to the 65 items were summed to provide a total score for analysis. Patient satisfaction with health services was measured with a self-report questionnaire (Tyrer and Remington, 1979) which comprised nine satisfaction items rated on a four-point scale (1–4). Satisfaction was summarized by adding all response values. A high score on either variable indicates poor outcome, that is, more psychopathology or greater dissatisfaction with health services.

Original analyses (Burns et al., 1999) were in principle by intention-to-treat, but patients with missing outcome data were excluded (Hollis and Campbell, 1999). The primary analysis compared mean outcomes between the two groups using a t-test. Secondary analyses adjusted for baseline variables using analysis of covariance (ANCOVA). Subjects with missing baseline values were included in these analyses by the missing indicator method (see below). Intervention groups did not differ significantly for any outcome at year 2. When outcome variables were adjusted for socio-demographic characteristics and other baseline variables, results were similar to unadjusted analyses (Burns et al., 1999).

Methods

Models

We defined the following sequence of models, some of which apply to all subjects with the relevant outcome (full population), and others of which apply only to that subgroup of the full population which also has the corresponding baseline measured (reduced population). Formal model definitions are given in Boxes 1 and 2.

- Model 1: an unadjusted intervention effect was estimated by direct comparison of the two groups using the t-test. This analysis was applied to the *full population*, but also to the *reduced population* to permit comparison with subsequent analyses.
- Model 2: analysis of covariance (ANCOVA) to adjust for centre.
- Model 3: ANCOVA to adjust for centre and the corresponding baseline variable. This should improve the precision with which the intervention effect is estimated, but must be applied to the *reduced population*.
- Model 4: ANCOVA to adjust for centre and baseline, using the missing indicator method (*full population*). Missing baseline values are filled in with an arbitrary value (e.g. 0), and a dummy variable is created to indicate missing baseline outcome. The ANCOVA model then contains randomised group, centre, baseline, and the dummy variable. This method is not recommended for observational data because it may lead to inadequate control of confounding (Greenland and

Finkle, 1995) but this criticism does not apply in randomized trials.

- Model 5: random effects model 1 (REM-1) for outcomes at baseline and year 2, adjusting for centre. The variables used are the same as in model 3, but the random effects framework allows use of the *full population* and additionally includes all subjects with baselines, whether or not they have final outcomes. In the absence of missing baselines, it may be shown that REM-1 and ANCOVA give identical estimates of the intervention effect, because the likelihood for REM-1 factorizes as the likelihood of the baseline data multiplied by the ANCOVA likelihood.
- Model 6: random effects model 2 (REM-2) for outcomes at baseline, year 1 and year 2, adjusting for centre (*full population*). This model implicitly uses data from baseline and year 1 to give information about missing year-2 outcomes.
- Model 7: multivariate random effects model (MV-REM) for all three outcomes – psychopathology, satisfaction and days in hospital – at baseline, year 1 and year 2, adjusting for centre (*full population*). Days in hospital were log-transformed (after adding one day) to reduce skewness. This model makes use of recorded days in hospital for a number of individuals who were not interviewed.

Estimation

We fitted each of these models using a likelihood-based approach to the observed data. Because the missing data mechanism is not modelled, this approach makes a missing-at-random (MAR) assumption that the probability of the data point being missing is independent of the unobserved data, *conditional on all observed data in the model*. Specifically, estimates from model 1 assume that, in each randomized group, individuals with missing and non-missing outcomes at year 2 have the same distributions of true outcomes. In models 2–4, individuals with missing and non-missing outcomes may differ, so long as these differences are explained by other variables Z , such as centre and baseline (in model 3). In the random effects models 5–7, the MAR assumptions are complicated because individuals with missing baselines may be observed at 2 years ('non-monotonic' missingness; Robins and Gill, 1997). However, sufficient conditions for MAR to hold may again be

Box 1: notation

Y_{ij}	Outcome variable (CPRS or satisfaction) for person i at occasion j ($j = 0, 1, 2$).
$CPRS_{ij}, SAT_{ij}, HOS_{ij}$	Specific outcome variables (symptom rating on CPRS, satisfaction, log(days in hospital+1)) for person i at occasion j .
Y_{ij}^*	Y_{ij} with missing values replaced by 0.
M_{ij}	Indicator of missing Y_{ij} (1 if Y_{ij} is missing, 0 otherwise).
$RAND_i$	Randomization status: 0 = standard case management, 1 = intensive case management
$CENTRE_i$	Centre from which patient was drawn (0, 1, 2, 3).
$C1_i, C2_i, C3_i$	Indicators of centre 1, 2, 3. $C1_i = 1$ if patient i is in centre 1, 0 otherwise.
e_{ij}	Residual from model.

Box 2: model definitions*Simple models*

1. $Y_{i2} = \alpha + \beta RAND_i + e_{i2}$
2. $Y_{i2} = \alpha + \beta RAND_i + \gamma_1 C1_i + \gamma_2 C2_i + \gamma_3 C3_i + e_{i2}$
3. $Y_{i2} = \alpha + \beta RAND_i + \gamma_1 C1_i + \gamma_2 C2_i + \gamma_3 C3_i + \delta Y_{i0} + e_{i2}$
4. $Y_{i2} = \alpha + \beta RAND_i + \gamma_1 C1_i + \gamma_2 C2_i + \gamma_3 C3_i + \delta_1 Y_{i0}^* + \delta_2 M_{i0} + e_{i2}$

Random effects models

5. $Y_{ij} = \alpha_j + \beta_j RAND_i + \gamma_{j1} C1_i + \gamma_{j2} C2_i + \gamma_{j3} C3_i + e_{ij}$ $j=0,2$
 α_j = control mean at year j , β_j = treatment effect at year j ($\beta_0 = 0$). e_{i0}, e_{i2} have a bivariate normal distribution.
6. As REM-1, but $j = 0,1,2$, so e_{i0}, e_{i1}, e_{i2} have a trivariate normal distribution.
7. $CPRS_{ij} = \alpha_j^C + \alpha_j^C RAND_i + \gamma_{j1}^C C1_i + \gamma_{j2}^C C2_i + \gamma_{j3}^C C3_i + e_{ij}^C$ $j=0,1,2$
 $SAT_{ij} = \alpha_j^S + \beta_j^S RAND_i + \gamma_{j1}^S C1_i + \gamma_{j2}^S C2_i + \gamma_{j3}^S C3_i + e_{ij}^S$ $j=0,1,2$
 $HOS_{ij} = \alpha_j^H + \beta_j^H RAND_i + \gamma_{j1}^H C1_i + \gamma_{j2}^H C2_i + \gamma_{j3}^H C3_i + e_{ij}^H$ $j=0,2$
 $\beta_0^C = \beta_0^S = \beta_0^H = 0$. $e_{i0}^C, e_{i1}^C, e_{i2}^C, e_{i0}^S, e_{i1}^S, e_{i2}^S, e_{i0}^H, e_{i2}^H$ have a multivariate normal distribution.

expressed in terms of 'explaining variables' Z : in model 5, Z comprises centre, missingness of baseline, and baseline (if observed); in model 6, missingness of year 1 observation and year 1 value (if observed) join Z ; and in model 7, concurrent outcomes join Z . Similar assumptions are also required for missingness at baseline and year 1. These assumptions become progressively more plausible as more variables and time-points are included in the model.

In all random effects models, the random terms are assumed to be normally distributed with mean zero. In Model 6 (REM-2), an unstructured covariance matrix was used for the deviations at times 0, 1 and 2. An alternative formulation would be a random slopes model, but this gave a worse fit to the data and, by overestimating the variance at year 2, it overestimated

standard errors for the year 2 intervention effect. Similarly, in Model 7 (MV-REM), an unstructured covariance matrix was used. The models could therefore be described as multivariate ANOVA adapted to handle missing data. Models including both year 1 and year 2 data allowed separate intervention effects at the two time points.

T-tests and ANCOVAs were performed using Stata (Stata Corp, 1999). Random effects analyses were executed in MLwiN (Rasbash, Browne and Goldstein, 1999), using the restricted iterative generalized least squares algorithm to provide REML parameter estimates; in SAS using PROC MIXED (Littell et al., 1996); and in the structural equation modelling software Mplus (Muthén et al., 1998) using direct maximum likelihood methods (Arbuckle, 1996).

Artificially missing data

To explore the performance of our sequence of models under known missingness mechanisms, we refitted all models for satisfaction after artificially deleting some data, and compared the results with those obtained for the full, observed data set. Specifically, satisfaction at year 2 was deleted in the standard arm only whenever it lay above the upper quartile. This is an informative missing mechanism.

Results*Trial description*

The UK700 trial recruited 708 subjects, of which 353 were allocated to intensive case management and 355

to standard case management. Table 1 shows mean outcomes over time for the two randomized groups. Correlations between the outcomes were relatively weak, ranging from 0.25 to 0.59 between the same outcome at different times and from 0.02 to 0.34 between different outcomes at the same time (Table 2).

Pattern of missing data

The extent of missingness was comparable between randomized groups for each outcome variable at baseline. There was a significantly higher proportion of subjects not interviewed in the standard case management group (18% versus 10% at year 2, $p < 0.01$). The standard case management group therefore displayed a higher proportion of missing

Table 1. Clinical characteristics of participants at baseline and follow-up

Outcome		Intensive case management (n = 353)		Standard case management (n = 355)	
		Mean (standard deviation)	n	Mean (standard deviation)	n
Baseline	CPRS	19.7 (13.0)	350	17.9 (12.3)	355
	Satisfaction	19.0 (4.8)	281	18.8 (4.8)	290
	Days in hospital	105.7 (111.0)	352	111.9 (114.2)	355
Year 1	CPRS	16.4 (12.1)	304	18.1 (14.1)	289
	Satisfaction	17.0 (4.4)	263	17.5 (5.0)	248
Year 2	CPRS	18.5 (13.8)	311	18.1 (13.9)	284
	Satisfaction	16.7 (4.9)	258	17.1 (4.7)	232
	Days in hospital	73.5 (124.2)	338	73.1 (111.2)	341

Table 2. Covariance matrix of outcomes used (correlations above diagonal, variances on diagonal, covariances below diagonal)*

		CPRS			Satisfaction			Days in hospital	
		Baseline	Year 1	Year 2	Baseline	Year 1	Year 2	Baseline	Year 2
CPRS	Baseline	163.54	0.36	0.40	0.21	0.15	0.15	0.02	0.10
	Year 1	60.12	171.43	0.59	0.06	0.26	0.20	0.10	0.12
	Year 2	73.16	108.79	200.45	0.11	0.14	0.34	0.10	0.06
Satisfaction	Baseline	13.10	3.78	7.25	23.25	0.31	0.29	0.12	0.14
	Year 1	9.35	16.43	9.48	7.33	23.94	0.47	0.13	0.19
	Year 2	8.85	12.28	22.82	6.68	10.86	22.41	0.19	0.30
Days in hospital	Baseline	0.21	1.39	1.44	0.62	0.68	0.96	1.14	0.25
	Year 2	2.75	3.51	1.83	1.41	1.96	3.02	0.58	4.67

* Analyses based on 354 subjects with complete data. Days in hospital were log-transformed after adding one day.

values at years 1 and 2 for the outcome variables collected by interview, but not for days in hospital (Table 3); 112 participants had neither CPRS nor satisfaction reported at year 2 (98 not interviewed and a further 14 interviewed but with both variables missing); of these, days in hospital were recorded for 83 (74%).

The longitudinal patterns of missing data are shown in Table 4. Missingness on one occasion increased the chance of missingness on a subsequent occasion. For example, for satisfaction, missing data at baseline increased the subsequent chance of missing data to 53% as compared to 38% for those with recorded values of baseline satisfaction. For CPRS, missing data at year 1 increased the chance of missing data at year 2 to 57% as compared to 8% for those with recorded values at year 1.

Only year 1 satisfaction differed significantly between those interviewed and non-interviewed at year 2 (Table 5). There was no significant difference between the number of days spent in hospital at year 2 between interviewed and non-interviewed subjects for whom hospitalization data were available (mean difference 0.66, standard error 14.97). The proportion of subjects not interviewed at year 2 varied significantly by centre ($p = 0.01$) but not by sex, ethnicity, diagnosis, or father's occupation at birth.

Simple analyses of intervention effect

For CPRS, the t-test (Model 1) showed a small and non-significant harmful effect of intensive case management (Table 6). Adjustment for centre (Model 2) produced a moderate change in favour of standard case management. Adjustment for baseline (Model 3)

Table 4. Longitudinal patterns of missingness

Pattern*	CPRS	Satisfaction
OOO	546	355
OOX	45	70
OXO	49	55
OXX	65	91
XOO	0	65
XOX	2	21
XXO	0	15
XXX	1	36

* O denotes an observed value, X denotes a missing value. For example, OXX denotes only observed data at baseline, and OXO denotes data missing at year 1 but observed at baseline and year 2.

changed the point estimate to favour intensive case management, because of a chance baseline imbalance in CPRS. The standard error decreased by 8%, but the estimated effect remained non-significant. All 595 subjects in this analysis had complete data on baseline measures, so model 4 agreed exactly with model 3 and is not shown.

For satisfaction, all estimated effects were small and non-significantly favoured intensive case management. Excluding the 80 subjects with missing baseline satisfaction changed the estimated mean difference in the t-test from -0.33 to -0.16 and increased the standard error by 9%. Compared with adjustment for centre only (Model 2), therefore, additional adjustment for baseline (Model 3) increased the standard error by 6% in the simple analysis but decreased the standard error by 3% using the missing indicator method (Model 4).

Table 3. Number of missing values (%) by treatment group at baseline, year 1, and year 2

Outcome	Baseline		Year 1		Year 2	
	Intensive	Standard	Intensive	Standard	Intensive	Standard
CPRS	3 (1%)	0 (0%)	49 (14%)	66 (19%)	42 (12%)	71 (20%)
Satisfaction	72 (20%)	65 (18%)	90 (26%)	107 (30%)	95 (27%)	123 (35%)
Days in hospital	1 (0%)	0 (0%)	n.a.	n.a.	15 (4%)	14 (4%)

Table 5. Difference in means between subjects interviewed at year 2 and those not interviewed

Outcome	Baseline			Year 1		
	Mean difference* (standard error)	P	n	Mean difference* (standard error)	P	n
CPRS	0.64 (1.39)	0.64	705	0.08 (2.20)	0.96	593
Satisfaction	−0.40 (0.62)	0.51	571	−1.75 (0.90)	0.05	511
Days in hospital**	−8.43 (12.31)	0.52	707	n.a.	n.a.	n.a.

* Interviewed – not interviewed

** Untransformed

n.a. = not applicable

Random effects analyses of intervention effect

Results of the first random effects model (REM-1) agreed closely with the missing indicator method (Table 6 and Table 7). Adding year 1 measurements using model REM-2 decreased the standard errors of the coefficients by about 1%. Adding concurrent outcomes using model MV-REM produced a further reduction of 1% to 2% in the standard errors. The estimates of the intervention effect did not change substantially from the missing indicator method.

The results of these analyses are summarized in Figures 1 to 3.

Artificially missing data

When data were artificially not missing at random, with missingness depending on the actual year 2 outcome, all analyses were substantially biased. However, the bias progressively decreased as extra information was included in the model (Figure 4).

Discussion*Non-response bias*

In the UK700 trial, the proportion of patients interviewed at the final follow-up (year 2) differed significantly between intervention groups. This led to concern that simple analyses might be biased due to non-response. On the other hand, comparisons of baseline data, year 1 data, and hospitalization data between patients who were and were not interviewed at year 2 did not show significant differences, except for a borderline significant result for year 1 satisfaction scores. The latter results suggested that any non-response bias is likely to be small.

Our sequence of models aimed to shed further light on this question. The rationale for using more complex models was that, by including more baseline covariates and follow-up measurements, the data were more likely to approach MAR. For example, if patients who

Table 6. Analyses of intervention effect on CPRS at year 2

Model	Intervention effect		Records used		
	Estimate (Intensive – Standard)	Standard Error	Year 0	Year 1	Year 2
1. t-test	0.44	1.14	–	–	595
2. Adjusted for centre	0.69	1.12	–	–	595
3. Adjusted for centre + baseline ¹	−0.39	1.03	595	–	595
5. Random effects model 1	−0.39	1.03	705	–	595
6. Random effects model 2	−0.40	1.02	705	593	595
7. Multivariate random effects model	−0.31	1.01	705	593	595

¹ Models 3 and 4 give the same results since there are no missing baselines in subjects with non-missing outcomes.

Table 7. Analyses of intervention effect on satisfaction at year 2

Model	Population	Intervention effect		Records used		
		Estimate (intensive – standard)	Standard error	Year 0	Year 1	Year 2
1. t-test	Full	-0.33	0.43	-	-	490
1. t-test	Reduced	-0.16	0.48	-	-	410
2. Adjusted for centre	Full	-0.26	0.43	-	-	490
3. Adjusted for centre + baseline	Reduced	-0.26	0.45	410	-	410
4. Adjusted for centre + baseline (missing indicator)	Full	-0.35	0.41	490	-	490
5. Random effects model 1	Full+ ¹	-0.38	0.41	571	-	490
6. Random effects model 2	Full+ ¹	-0.35	0.41	571	511	490
7. Multivariate random effects model	Full+ ¹	-0.41	0.40	571	511	490

¹ Also includes patients with missing year 2 satisfaction but observed baseline or year 1 data.

Figure 1. Estimates of year 2 intervention effect on CPRS score. Models are 1 (t-test), 2 (analysis of covariance adjusting for centre), 5 (random effects model 1 including centre and baseline), 6 (random effects model 2 including centre, baseline and intermediate outcome) and 7 (multivariate random effects model including centre, intermediate outcomes and concurrent outcomes). All analyses use the full population of 595 patients at year 2.

were not interviewed at year 2 differed systematically in satisfaction from those who were interviewed, then one might expect also to see a difference in satisfaction measured at baseline, in satisfaction measured at year 1, and potentially in hospitalization over the 2 years of the study. Inclusion of any of these variables in the analysis model should reduce non-response bias, so the estimated intervention effect would be expected to change steadily as more complex models were fitted. In

the absence of non-response bias, changes in the estimated intervention effect are more likely to be small and unsystematic.

Figures 1 and 2 summarize the estimated intervention effects in the UK700 data for Models 1, 2, 5, 6 and 7. As more information is included, estimates of intervention effect change only little, and the changes are not in a consistent direction. It therefore seems reasonable to conclude that non-response bias was small.

Figure 2. Estimates of year 2 intervention effect on satisfaction. Models are as in Figure 1. All analyses use the full population of 490 patients at year 2.

Figure 3. Standard errors of year 2 intervention effect on CPRS and satisfaction. Models are 1 (t-test), 2 (analysis of covariance adjusting for centre), 3 (analysis of covariance adjusting for centre and baseline), 4 (analysis of covariance adjusting for centre and baseline using missing indicator method), 5 (random effects model 1 including centre and baseline), 6 (random effects model 2 including centre, baseline and intermediate outcome) and 7 (multivariate random effects model including centre, intermediate outcomes and concurrent outcomes).

Figure 4. Models as in Figure 1 applied to UK700 trial data, but with extra missing outcomes artificially created by an informative missing mechanism.

To determine whether this strategy is able to detect non-response bias where it does exist, we consider Figure 4. Although all estimates are grossly biased, there is a consistent trend, with the point estimates from MV-REM and the simple *t*-test differing by 1.5 standard errors. This trend therefore detects non-response bias but is unable to suggest what the true effect size is. With a less marked non-response bias, detection would clearly be difficult. The correlations of the outcome with the baseline, intermediate and concurrent outcomes in the UK700 trial were only around 0.3–0.4 (Table 2): the strategy will work better in trials where these correlations are larger.

Increasing precision

Intuitively, one would expect that including more data in the analysis would increase the precision of the estimated intervention effect. Indeed, in the data analysed here, adjustment for baseline increased precision (Model 4 versus Models 1 and 2), but excluding subjects with missing baseline decreased precision for the satisfaction outcome (Model 3 versus Models 1 and 2, Figure 3). Use of random effects models had little effect on standard errors, probably because of the relatively small correlations between occasions and between outcomes. Of the analyses that came near to maximum precision, the simplest was ANCOVA with missing baselines dealt with by the missing indicator method (Model 4).

Modelling issues

The simpler models can be fitted using widely used software such as Stata or SPSS. The random effects models, however, require more specialized software such as MLwiN (Rasbash, Browne and Goldstein, 1999), SAS proc MIXED (Littell et al., 1996) or MIXREG (Hedeker and Gibbons, 1996). They may also be fitted using structural equation software (Muthén et al., 1998), provided that estimation is based on the raw data (Arbuckle, 1996) and not on covariance matrices as are commonly used. The structural equation modelling tradition pays much attention to possible differences in variance-covariance matrices between groups, and this could also be done in the other software packages. It would however be important to constrain baseline variances to be equal between randomized groups.

For this trial, it was important to model the intervention effect flexibly over time. This will be true in many trials in which the aim is to assess the longer term effects of interventions. We used unstructured covariance matrices to model individual deviations from group averages but many other formulations are consistent with flexibly modelling the intervention effect. For example, a random slopes covariance structure can be combined with a saturated model for within-group means: in this, the means could follow an arbitrary shape, but individual regressions would be constrained to deviate from the overall means by a

linear function of time, and individual observations would deviate from the individual regression by uncorrelated or autocorrelated error terms. In other trials, it will be reasonable to make stronger assumptions about the intervention effect over time – for example, that it is proportional to time. In this case, a common analysis uses summary statistics (Frison and Pocock, 1997), but if random effects models are used then adjustment for baseline and inclusion of concurrent outcomes would be informative supplementary analyses.

Other methods

Other techniques for dealing with missing data include multiple imputation (Rubin, 1996), which essentially makes the same MAR assumptions as modelling using maximum likelihood. Our modelling strategy could also be used in this context, giving a sequence of increasingly complex imputation models. Imputed data sets would be analysed, for example by ANCOVA or multivariate analysis of variance, and the results using the different imputation models would be compared.

Conclusions

The analyses performed here reach the same conclusions as the original UK700 analyses: there are no convincing differences in outcomes for severely psychotic patients under standard or intensive case management two years after randomization to intervention. These results strengthen the original findings by demonstrating that models that include more data, and are less susceptible to bias, continue to show the same lack of effect of intervention. The mechanism for non-response in this trial of severely mentally ill patients living in the community appears to be that missingness depends only on randomized group and centre, and both simple and complex analyses appear to be largely free of non-response bias. However, it is never possible to be sure that non-response bias is absent. Further work is required to investigate the sensitivity of our proposed method to realistic degrees of non-response bias and to see how this varies with the correlation between outcomes.

Several features of a trial's results may suggest the possibility of non-response bias: differences in response rates between randomized groups; differences in observed baselines between randomized groups in those with non-missing final outcomes; and differences in observed baseline or intermediate outcomes between subjects with missing and non-missing final

outcomes. If none of these features is present then simple methods of analysis may be adequate. We recommend ANCOVA because of its greater precision; individuals with missing baseline values should not be excluded, and the missing indicator method is the simplest way to include them.

On the other hand, in the presence of any of these features suggestive of non-response bias, fitting the more complex models described here is valuable. If estimated intervention effects change consistently as more variables are included in the model, then non-response bias is probably present, and even the fullest model may have considerable residual bias. Formal sensitivity analysis for informative missingness would then be useful (Molenberghs, Kenward and Goetghebeur, 2001).

Acknowledgements

We thank the UK700 investigators for permission to use the UK700 trial data.

References

- Anderson TW. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J Am Stat Assoc* 1957; 52: 200–3.
- Arbuckle JL. Full information estimation in the presence of incomplete data. In Marcoulides GA, Schumacker RE (eds) *Advanced Structural Equation Modeling: Issues and Techniques*. Mahwah, NJ: Lawrence Erlbaum Publishing, 1996.
- Asberg M, Montgomery S, Perris C, Schalling D, Sedvall G. A comprehensive psychopathological rating scale. *Acta Psych Scand (Supplement)* 1978; 271: 5–27.
- Burns T, Creed F, Fahy T, Thompson S, Tyrer P, White I. Intensive versus standard case management for severe psychotic illness: a randomised trial. *Lancet* 1999; 353: 2185–9.
- Cook N. Restricted Maximum Likelihood. In Armitage P, Colton T (eds) *Encyclopedia of Biostatistics*. Chichester: John Wiley & Sons, 1998.
- Everitt BS. Analysis of longitudinal data: Beyond MANOVA. *Br J Psych* 1998; 172: 7–10.
- Frison L, Pocock SJ. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Stat Med* 1997; 16: 2855–72.
- Gibbons RD, Hedeker D, Elkin I, Waternaux C, Kraemer HC, Greenhouse JB, Shea T, Imber SD, Sotsky SM, Watkins JT. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Arch Gen Psych* 1993; 50: 739–50.
- Goldstein H. *Multilevel Statistical Models*, 2 edn. London: Edward Arnold, 1995.
- Greenland S, Finkle W. A critical look at methods for

- handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 1995; 142:1255–64.
- Hedeker D, Gibbons RD. MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Comput Meth Programs Biomed* 1996; 49: 229–52.
- Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *Br Med J* 1999; 319: 670–4.
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38: 963–74.
- Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS system for mixed models. SAS institute: Cary, NC, 1996.
- Little R, Rubin D. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 1987.
- Molenberghs G, Goetghebeur EJT, Lipsitz SR, Kenward MG, Lesaffre E, Michiels B. Missing data perspectives of the fluvoxamine data set: a review. *Stat Med* 1999; 18: 2449–64.
- Molenberghs G, Kenward MG, Goetghebeur E. Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Applied Statistics* 2001; 50: 15–29.
- Murray GD, Findlay JG. Correcting for the bias caused by drop-outs in hypertension trials. *Stat Med* 1988; 7: 941–6.
- Muthén LK, Muthén BO. *Mplus User's Guide*. Los Angeles CA: Muthén & Muthén, 1998.
- Rasbash J, Browne W, Goldstein H. *A User's Guide to MLwiN*. London: Institute of Education, 1999.
- Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. *Stat Med* 1997; 16: 39–56.
- Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; 91: 473–89.
- Schafer J. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.
- Shadish WR, Hu X, Glaser RR, Kownacki RJ, Wong T. A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychol Methods* 1998; 3: 3–22.
- Stat Corp. *Intercooled STATA 7.0 for Windows*. College Station, Texas: Stata Corporation, 1999.
- Tyrer P, Remington M. Controlled comparison of day-hospital and outpatient treatment for neurotic disorders. *Lancet* 1979; (i): 1014–16.
- Verbeke G, Molenberghs G (eds). *Linear Mixed Models in Practice*. New York: Springer-Verlag, 1997.

Correspondence: Ian White, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK.

Telephone (+44) (0)1223-330399.

Fax (+44) (0)1223-330388.

Email ian.white@mrc-bsu.cam.ac.uk.