ELSEVIER

Journal of Clinical Epidemiology 65 (2012) 90−99

Journal of Clinical Epidemiology

Telephone interviews can be used to collect follow-up data subsequent to no response to postal questionnaires in clinical trials

Ranjit Lall[a,*], Dipesh Mistry[a], Chris Bridle[a], Sallie E. Lamb[a,b]

[a]*Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Gibbett Hill Road, Coventry CV4 7AL, United Kingdom*
[b]*Kadoorie Critical Care Research Centre, University of Oxford, Oxford, United Kingdom*

Accepted 6 April 2011; Published online 9 August 2011

Abstract

Objective: Follow-up data were collected using postal questionnaires and if participants did not respond, then data was collected using telephone interviews. The objectives of this study were to examine, for the two methods, how respondents differed in characteristics and whether the observed treatment difference varied.

Study Design and Setting: A large clinical trial of lower back pain.

Results: About 60% (98/163) of the nonresponders to postal questionnaire provided data by telephone, which increased the overall response rate by 14% (from 71% to 85%). A consistent treatment difference was found across the methods for the outcome measures at 12 months, implying that the observed treatment effect had not been modified. There were some differences between the participants: responders of postal questionnaire were older, likely to be female, white (ethnic origin), not working, with less disability of back pain, compared with those who responded by a telephone interview. At 12 months, there was greater improvement in back pain, disability, and general health for those who responded by postal questionnaires.

Conclusion: Researchers should consider the use of more than one method of collecting data as this increases response rate, participant representativeness, and enhances precision of effect estimates.

Keywords: Clinical trial; Treatment; Postal; Questionnaire; Telephone; Follow-up

1. Introduction

Several methods for collecting follow-up data from participants who take part in research studies have been cited in the literature and these include the use of the postal questionnaires, face-to-face interviews, telephone interviews, and use of the Internet [1−3]. The postal questionnaire is the most frequently used and is considered to be the most cost effective but is often associated with the lowest response rate [4]. Poor response to questionnaires is known to reduce the statistical power of the study as the ''effective'' sample size is reduced. It can also introduce bias if the nonresponders are systematically different on outcomes of interest to those who respond to the questionnaires.

Much of the literature focuses on assessing ways of improving response to postal questionnaires at follow-up (e.g., giving incentives, use of shorter questionnaires, or use of reminders by telephone to return questionnaires) [5,6]. The use of an additional method, such as a telephone interview, as a means of collecting follow-up data from mailed nonresponders has been reported in a limited number of social and health surveys [7−10]. However, there are no studies that report the use of two or more methods of collecting follow-up data in a randomized clinical trial setting, where the comparison of interventions or treatments is of focal interest.

This observational study uses data obtained from a large clinical trial of back pain—the Back Skills Training Trial (BeST) [11]—where participants were randomized to one of two complex interventions. The use of the postal questionnaire was the primary method of data collection at follow-up, and if participants had not responded to mailed questionnaires, then attempts were made to capture data on a shorter version of the paper clinical forms through telephone interviews.

The primary aim of this article was to assess whether the estimated intervention effects differed for the two methods of data collection: postal questionnaires or telephone interviews, which were subsequently documented on paper questionnaires. A secondary aim was to explore how those who responded to the follow-up postal questionnaire differed to those who responded to a telephone interview.

* Corresponding author. Tel.: +44-247-657-4649; fax: +44-247-615-0549.

E-mail address: r.lall@warwick.ac.uk (R. Lall).

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.04.011

What is new?

Key findings?

The key findings in this study were that participants who responded to postal questionnaire differed in some of their demographic characteristics and baseline outcomes compared with those who responded by telephone interviews. These differences may be a potential source of bias, which occur because the methods of data collection differ in their implementation. However, if any degree of bias was present, this was consistent across the treatment arms. The implication of this was that an overall treatment difference for an outcome could be obtained irrespective of the way the data have been collected.

What this adds to what was known?

In clinical trials, where the comparison of interventions is of focal interest, follow-up data are often collected using a single method, for example, postal questionnaires. In this article, we have illustrated that collecting data on respondents of postal questionnaires supplemented with collecting data on nonrespondents by telephone can increase response and increase generalizability.

What is the implication, what should change now?

The implication of this article is: telephone interviews should be considered as an effective way in reducing loss to follow-up of participants in clinical trials; researchers should consider additional methods of collecting data from participants who are not responding to initial requests. However, if different methods are used to collect data in a clinical trial, then the data need to be carefully scrutinized for any biases and how these may affect the overall estimate of treatment effect.

For both these aims, the following were evaluated: (1) demography and baseline outcomes; (2) response rates; (3) missing data; (4) internal consistency of each outcome, and (5) the 12 month outcome measures.

2. Subjects and methods

2.1. Design and participants

The BeST [11,12] was a two-armed pragmatic multicentered randomized controlled trial. The main aim of this trial was to estimate the clinical effectiveness of two complex interventions: active management (AM) vs. AM plus a cognitive behavioral approach (CBA) (AM + CBA) for subacute and chronic low back pain. AM is the standard intervention for back pain and consists of an advice session supplemented by the Back Book [13]. The CBA intervention details a cognitive-based program, which covers several areas, thought to be fundamental for managing back pain—pain assessment and coping strategies, goal setting, and exercise and its benefits [14]. This program was delivered in group sessions over a 6-week period. All eligible participants, who satisfied the study criteria, were required to complete a baseline questionnaire, after which they were randomly assigned to one of the two therapy groups in a ratio of 1:2 (AM: AM + CBA). An independent telephone randomization service (based at the Medical Research Council Clinical Trials Unit in London) was used and randomization was generated using stratified blocks, where the strata were the center and severity of back pain (moderate, severe/very severe). Follow-up data were collected 12 months after randomization using postal questionnaires. Nonrespondents of the initial self-reported questionnaire were sent another questionnaire with a reminder letter after 2 weeks of the first questionnaire being mailed out. If the latter still had not been returned after another 2 weeks (i.e., after 4 weeks of the first questionnaire being mailed out), participants were telephoned to remind them to fill in the questionnaire or if they had misplaced their copy, then arrangements were made for another questionnaire to be sent out. If the questionnaire remained unreturned after a further 2 weeks (i.e., 6 weeks of the first questionnaire going out), a telephone call was made to the participant to ask if it was possible to collect a limited set of data over the telephone.

2.2. Telephone interview process

Telephone interviews were conducted by three health professionals (one research nurse and two research physiotherapists) who had initially recruited participants into the trial. The content of the interviews were standardized, and the health professionals were trained in how to administer the telephone interviews. The interview required the completion of a shorter version of the postal questionnaire. The postal questionnaire comprised seven validated instruments and three of these were selected for the telephone interview and completed in full. The order and layout of the questions was the same as presented in the postal questionnaire. Interviewers were instructed to read the question and response categories exactly as written in the paper questionnaire. They then recorded the response provided by the participant. If the participant requested further information or explanation, the question was read out again but no additional information was offered. Telephone interviews lasted an average of 5 minutes.

2.3. Outcome measures

The outcome measures used to assess treatment efficacy in both the postal and telephone questionnaires were the modified Von Korff scale [15], the Short Form 12 (SF-12) [16], and the Euro-QoL (EQ-5D) [17]. The modified Von

Korff scale was one of the primary outcome measures in the BeST [11] and was used to assess back pain and disability in the last 4 weeks. It comprised of six questions, each scored on a scale from 0 (no pain/disability) to 10 (worst pain/disability). The first three questions related to disability and asked about how back pain interfered with (1) daily activity; (2) recreation; and (3) ability to work. The last three questions related to pain and assessed (1) worst pain; (2) average pain; and (3) rating of back pain today. Each subscale (pain and disability) was transformed into a score ranging from 0 (no pain/disability) to 100 (worst pain/disability). In addition to this, we also collected data on quality of life using the SF-12 and general health status using the EQ-5D. The physical and mental health composite scores were computed from the SF-12 measure. These scores were obtained from 12 questions and ranged from 0 to 100, with a higher score indicating better physical/mental functioning. The EQ-5D is a standardized instrument for use as a measure of health outcome and a higher score signifies a better health-related quality of life. It is based on five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) and these were converted into a single summary score typically ranging from 0 (death) to 1 (perfect health state).

2.4. Statistical analysis

The data collected on the postal questionnaires and that provided by the telephone interviews as recorded on paper questionnaires were entered into an ACCESS 2003 (Microsoft) [18] database by one data entry clerk. A 100% validation check was carried out on a random sample of 10% of participants drawn from the database. All these data were checked against the paper questionnaires (both postal and those used to record the responses from the telephone interview) for correctness. Acceptable error rates for data entry (i.e., $<0.5\%$ as given in [19]) were obtained. All the data entry, cleaning, and validation processes were identical for both sources of data. The statistical analysis was carried out using STATA version 10 [20]. All *P*-values less than 0.05 were regarded as statistically significant.

2.4.1. Response rates at follow-up

Response rates were summarized for (1) each method of data collection within the two intervention arms and (2) methods of data collection alone. In the CONSORT diagram [21], the percentage of participants was based on total number randomized within each treatment arm. The chi-square test statistic was used to assess the association of response to questionnaire (postal and telephone) and intervention group.

2.4.2. Demographic characteristics and baseline measures

Participant characteristics and baseline measures were summarized using unadjusted summaries for (1) each intervention by the method of data collection and (2) the method of data collection alone. A linear regression model was fitted for each continuous demographic and baseline measure, with independent variables as treatment, method of data collection, and the interaction of treatment and method of data collection. The purpose of these linear regression models was to assess whether the difference in the interventions was consistent across the two methods of data collection (using the interaction term). In the case of a nonsignificant interaction term, two linear regression models were fitted to assess the difference in the methods of data collection: (1) with only the main effect of the method of data collection in the model and (2) with main effects of treatment and methods of data collection. The results were very similar for both of these models and therefore the difference in the response (and 95% confidence intervals) provided by the two data collection methods and their *P*-values were reported using the models where adjustments were made for the treatment effect. Categorical variables were fitted in a similar way using logistic regression models and variables with more than two categories were fitted using polytomous models. Ethnicity was found to be sparse in some of the categories, and the polytomous model allowing for the exact test was used to account of this. The adjusted odds ratios (and the 95% confidence intervals) were provided for each of these models, with the referent category as the last category on the scale. The *P*-values were obtained for these odds ratios.

2.4.3. Missing data at follow-up

The total number of questions for each outcome where participants had not provided a response were documented and summarized for (1) each type of data collection method within each of the treatment interventions and (2) for each method of data collection alone. The percentage of missing questions for a given scale was computed as follows:

$$\frac{\text{number of missing questions}}{(\text{number of responders}) \times (\text{number of questions presented})}$$

where the number of responders is provided in the CONSORT diagram. For a given outcome, any one participant could have more than one question unanswered of the total number of questions, which form the inventory score. Because of this dependence, we attempted to fit a two-level multihierarchical random effect logistic regression model, where the dependent variable was the missing/nonmissing item and the random effect variable was participant at level 2 with the individual questions nested at level 1. The fixed effects were treatment intervention and method of data collection together with the interaction term of these two fixed effects. It was not possible to computationally fit these models because of the small frequencies in some of the cells. Instead, we had to use the

ordinary exact logistic regression models, which assumed independence in the items and accounted for the small cell frequencies. Both the mental and physical component scores use all the 12 questions presented on the SF-12 scale.

2.4.4. Internal consistency

The Cronbach's alpha, which is a coefficient of internal consistency, was computed for each of the instruments (modified Von Kroff [disability and pain], SF-12, and the EQ-5D), for (1) both methods of data collection within each treatment and (2) for both methods of data collection alone. The internal consistency indices were computed using all the SF-12 items for the mental and physical component scores. The Cronbach's alpha indicates the degree to which individual questions group together to form combined scale indices. A low (<0.7) Cronbach's alpha indicated that the instrument was no longer measuring a single trait variable.

2.4.5. Outcome measures at follow-up

In the main effectiveness analysis [11], the focus was on the evaluation of a treatment difference where the outcome measures were analyzed as change from baseline (baseline−follow-up). Linear regression models were fitted with the independent variable as the treatment intervention, adjusted for the covariates gender, age of the participant, and baseline outcome and therefore no other adjustments for significant baseline variables was made. In the analysis presented in this article, the aim was to see if the methods of data collection had resulted in different estimates of the intervention effect, which was of ultimate interest. Thus, the latter regression models were fitted, with the inclusion of the method of data collection as a main effect term and the interaction term of treatment intervention and method of data collection. The interaction term in the model was used to assess whether the observed intervention difference were similar over the method of questionnaire administration. In the case of nonsignificant interaction terms, only the main effect model was fitted. This provided the adjusted means and mean difference (and 95% confidence intervals) for the two levels of the data collection variable, given that the treatment and other covariates were present.

3. Results

3.1. Response rates at follow-up

In total, 701 participants were randomized into the BeST trial, with 233 participants randomized to AM and 468 to AM + CBA (Fig. 1). This was the total planned sample size for this clinical trial. All randomized participants completed a baseline clinical record form providing some or all of their baseline data. Follow-up data at 12 months was available for 85.4% (598/701) of all patients randomized. The total response rate to the postal questionnaire at 12 months was 71.3% ($n = 500$). Of the nonresponders, 23.3% ($n = 163$) were subsequently contacted by telephone for follow-up data and 5.4% ($n = 38$) had withdrawn from the trial. Approximately 60% of these mailed nonresponders provided outcome data by

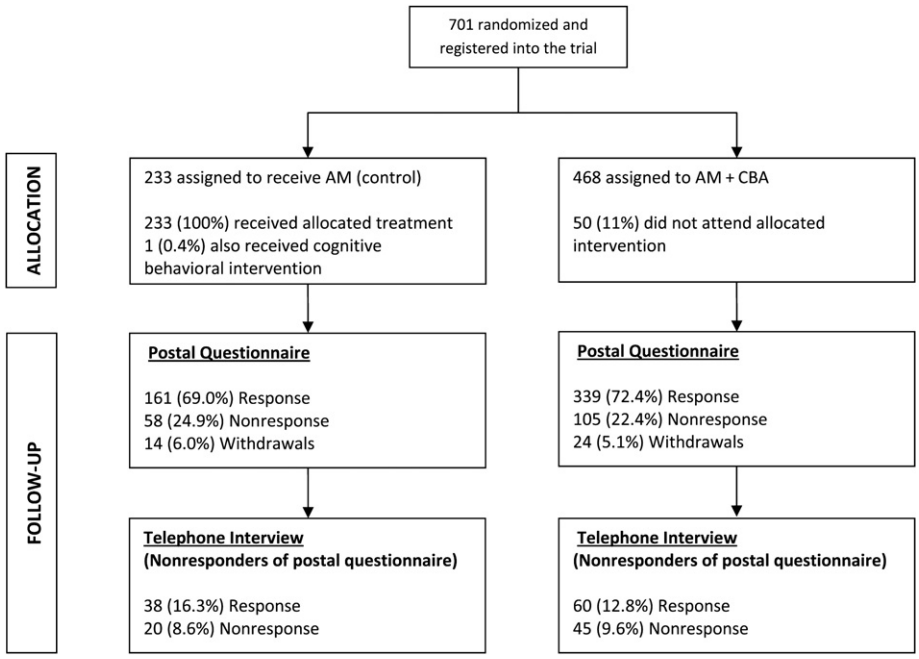

Fig. 1. Flow diagram showing the flow of participants for the BeST trial. *Abbreviations:* AM, active management; CBA, cognitive behavioural approach; BeST, Back Skills Training Trial.

telephone. This increased the overall response rate of the trial by 14% from 71% to 85%. Of those randomized to the AM arm, 16.3% ($n = 38$) participants provided data by telephone. Of those who were randomized to the AM + CBA arm, 12.8% ($n = 60$) subjects provided data by telephone. There was no evidence to suggest that there was an association between the number of participants responding (mailed and telephone) across the intervention arms and the method of data collection (Chi-square test statistic: $P = 0.21$).

3.2. Demographic characteristics and baseline measures

No significant interaction terms (treatment × method of data collection) in the linear regression models were apparent, implying that the demographic and baseline measures were consistent across the two methods of data collection (Tables 1 and 2). Omitting the nonsignificant interaction meant that we could assess the difference in participant characteristics for the methods of questionnaire administration. Postal questionnaire respondents were significantly older compared with those responding by telephone. There was some suggestion that males were more likely to respond to a telephone interview than females, and that females were more likely to respond to a postal questionnaire than males (P-value = 0.032). There were more participants of white ethnicity who responded to postal questionnaires than those of any other origin. There was significant evidence to suggest that participants not in employment were much more likely to respond to a postal questionnaire compared with those interviewed. All of the baselines outcome measures with the exception of the SF-12 physical score were found to be significantly different when comparing the two methods of data collection. Participants responding by telephone reported higher mean modified Von Korff scores and lower SF-12 mental and EQ-5D scores compared with those responding by post. This implied that participants who were interviewed were generally poorer in health (i.e., back pain, disability, mental state, and health-related quality of life) compared with those who were followed up by a postal questionnaire.

3.3. Missing data at follow-up

The proportion of missing questions were higher on the postal questionnaire compared with the questionnaire used to collect the interview data (Table 3). The interaction terms from the logistic regression models could not be fitted for the modified Von Korff scores because of the presence of ''no missing data'' for the telephone assessments. The interaction term in the models for the SF-12 and EQ-5D scores was found to be significant. The implication of this was that the degree of missingness between the treatment arms was not similar over the methods of data collection. However, this result was interpreted with caution, as the test of interaction terms in both model were underpowered and the findings may be spurious.

3.4. Internal consistency

The Cronbach's alpha values illustrate that the unidimensionality of each instrument is acceptable and that the measures are indeed assessing a single trait (Table 4). The EQ-5D measure collected over the telephone for those on the AM arm illustrated a lower Cronbach's alpha (0.57) compared with the value produced for those on the AM + CBA arm. On visual inspection, with the exception of the latter, differences in the Cronbach's alpha values for the two intervention arms were approximately the same across the two methods of data collection. This suggests that the unidimensionality of the measures assessed for the two interventions is not affected by the method of data collection. Also, on comparing the indices obtained for the postal and telephone methods alone, there appeared to be no reason to assume that unidimensionality was not consistent.

3.5. Follow-up at 12 months

For each outcome, the interaction term of treatment and method of data collection in the model was nonsignificant, suggesting that the estimate of the treatment difference (based on change from baseline to 12 month scores) had not been influenced by the method of data collection (see Table 5). Assessment of the main terms regression models provided evidence of a significant difference in the change scores for the participants who were interviewed and those who provide postal response for the modified Von Korff Disability and EQ-5D scores. Participants who responded by postal questionnaire had greater improvement in their back pain and disability and a smaller deterioration in their general health compared with those who had responded by a telephone interview.

4. Discussion

The total sample size for the BeST trial of 701 participants took account of 25% loss to follow-up. However, despite this, it is recognized that some of the analyses presented are underpowered, particularly the test of the interaction terms in the regression models. Thus, the significant interaction between treatment and method for missing items should be interpreted with caution. Also, when comparing the two methods of data collection, it is advisable to place less emphasis on the P-values and give more attention to the point estimates and 95% confidence intervals. With this in mind, there is suggestion from the results that the characteristics and outcomes of participants who respond by the two methods of data collection are somewhat

Table 1. Demographic characteristics at baseline

Demographic characteristics	AM[a]		AM + CBA[a]		Total[b]		
	Postal	Telephone interview	Postal	Telephone interview	Postal	Telephone interview	Postal vs. Telephone[c]
Age (yr)							
N (%)	161 (80.9)	38 (19.1)	339 (85.0)	59 (14.8)	500 (83.8)	97 (16.2)	
Mean	56.1	46.1	54.7	47.8	55.2 (14.1)	47.1 (14.2)	8.1**
95% CI	54.04–58.26	41.28–50.93	53.16–56.20	44.25–51.28	53.92–56.39	44.27–49.95	5.0–11.2
Missing, *n* (%)	0	0	0	1 (100.0)	0	1 (100.0)	
Sex							
Male, *n* (%)	61 (79.2)	16 (20.8)	131 (79.9)	33 (20.1)	192 (79.7)	49 (20.3)	0.6 (0.4–0.9)
Female, *n* (%)	99 (81.2)	22 (18.2)	208 (88.5)	27 (11.5)	307 (86.2)	49 (13.8)	
Missing, *n* (%)	1 (100.0)	0	0	0	1 (100.0)	0	
Ethnicity							
White, *n* (%)	145 (82.9)	30 (17.1)	305 (87.1)	45 (12.9)	450 (85.7)	75 (14.3)	0.3* (0.1–1.4)
Asian or Asian British, *n* (%)	2 (33.3)	4 (66.7)	13 (72.2)	5 (27.8)	15 (62.5)	9 (37.5)	1.2* (0.2–6.0)
Black or Black British, *n* (%)	2 (66.7)	1 (33.3)	3 (50.0)	3 (50.0)	5 (55.6)	4 (44.4)	1.6* (0.2–10.8)
Chinese or other, *n* (%)	3 (75.0)	1 (25.0)	3 (60.0)	2 (40.0)	6 (66.7)	3 (33.3)	
Missing, *n* (%)	9 (81.8)	2 (18.2)	15 (75.0)	5 (25.0)	24 (77.4)	7 (22.6)	
Left full-time education							
Age ≤16, *n* (%)	81 (80.2)	20 (19.8)	195 (86.7)	30 (13.3)	276 (84.7)	50 (15.3)	0.7 (0.4–1.2)
Age 17–19, *n* (%)	45 (83.3)	9 (16.7)	75 (86.2)	12 (13.8)	120 (85.1)	21 (14.9)	0.6 (0.3–1.2)
Age ≥20, *n* (%)	28 (77.8)	8 (22.2)	53 (79.1)	14 (20.9)	81 (78.6)	22 (21.4)	
Missing, *n* (%)	7 (87.5)	1 (12.5)	16 (80.0)	4 (20.0)	23 (82.1)	5 (17.9)	
Work							
No, *n* (%)	91 (88.3)	12 (11.7)	171 (89.1)	21 (10.9)	262 (88.8)	33 (11.2)	2.2* (1.4–3.4)
Yes, *n* (%)	69 (72.6)	26 (27.4)	167 (81.1)	39 (18.9)	236 (78.4)	65 (21.6)	
Missing, *n* (%)	1 (100.0)	0	1 (100.0)	0	1 (100.0)	0	
Frequency of LBP							
Everyday, *n* (%)	108 (79.4)	28 (20.6)	228 (84.1)	43 (15.9)	336 (82.6)	71 (17.4)	1.1 (0.5–2.4)
Between everyday & 3/4 of the days, *n* (%)	24 (92.3)	2 (8.7)	57 (89.1)	7 (10.9)	81 (90.0)	9 (10.0)	0.6 (0.2–1.5)
Between 3/4 and 1/2 of the days, *n* (%)	11 (73.3)	4 (26.7)	29 (87.9)	4 (12.1)	40 (83.3)	8 (16.7)	
Missing, *n* (%)	18 (81.8)	4 (18.2)	25 (80.6)	6 (19.4)	43 (81.1)	10 (18.9)	
Severity							
Not at all troublesome	0	0	0	0	0	0	
Slightly troublesome, *n* (%)	8 (72.7)	3 (27.3)	17 (77.3)	5 (22.7)	25 (75.8)	8 (24.2)	2.2 (0.7–8.1)
Moderately troublesome, *n* (%)	79 (87.8)	11 (12.2)	160 (88.4)	21 (11.6)	239 (88.2)	32 (11.8)	1.0 (0.4–2.7)
Very troublesome, *n* (%)	49 (73.1)	18 (26.9)	107 (81.1)	25 (15.9)	156 (78.4)	43 (21.6)	2.0 (0.8–5.5)
Extremely troublesome, *n* (%)	7 (77.8)	2 (22.2)	30 (90.9)	3 (9.1)	37 (88.1)	5 (11.9)	
Missing, *n* (%)	18 (81.8)	4 (18.2)	25 (80.6)	6 (19.4)	43 (81.1)	10 (18.9)	

Abbreviations: AM, active management; CBA, cognitive behavioural approach; CI, confidence interval; LBP, low back pain.

Note: Mean difference is for the ''age'' variable only, for the remaining variables the odds ratios (with the last category on the scale as referent) have been used to express a comparison between the methods of follow-up.

*Indicates $0.001 \leq$ *P*-value < 0.05; ** indicates *P*-value < 0.001: *P*-value is provided from the ''data collection'' main effect term in the linear regression model: response $= \beta_0 + \beta_1$ treatment $+ \beta_2$ data collection.

[a] Row percentages are based on the total respondents within each treatment arm and each row.

[b] Row percentages are based on the total respondents with each row.

[c] Mean difference was expressed for the age variable only; for the remaining variables the odds ratio was used for comparison of the methods of data collection.

different: those who respond by telephone tend to be younger than those who respond by post. Postal respondents are likely to be female, white in ethnic origin, and not in employment, with baseline outcomes, which signify better health status in terms of their back pain and general health compared with those giving a telephone interview. Even after the interventions, those who respond by a postal questionnaire show more improvement in the outcomes

Table 2. Outcome measures at baseline

	AM[a]		AM + CBA[a]		Total[b]		
Outcome measures	**Postal**	**Telephone**	**Postal**	**Telephone**	**Postal**	**Telephone**	**Difference (postal–telephone)**
Modified Von Korff Disability							
N (%)	158 (80.6)	38 (19.4)	329 (84.6)	60 (15.4)	487 (83.2)	98 (16.8)	
Mean	43.2	56.5	47.4	50.3	46.0	52.7	−6.8*
95% CI	39.66, 46.80	49.19, 63.79	44.78, 49.93	44.07, 56.48	43.92, 48.11	47.93, 57.44	−12.0, −1.7
Missing, *n* (%)	3 (100.0)	0	10 (100.0)	0	13 (100.0)	0	
Modified Von Korff Pain							
N (%)	160 (80.8)	38 (19.2)	336 (84.8)	60 (15.2)	496 (83.5)	98 (16.5)	
Mean	57.1	63.2	57.8	63.4	57.6	63.4	−5.8*
95% CI	54.23, 60.02	57.34, 69.15	55.75, 59.92	59.16, 67.73	55.92, 59.30	59.90, 66.83	−9.9, −1.7
Missing, *n* (%)	1 (100.0)	0	3 (100.0)	0	4 (100.0)	0	
SF-12 Mental score							
N (%)	149 (81.0)	35 (19.0)	309 (84.9)	55 (15.1)	458 (83.6)	90 (16.4)	
Mean	47.4	42.9	45.5	40.9	46.1	41.7	4.6**
95% CI	45.61, 49.14	39.65, 46.10	44.22, 46.77	37.82, 43.98	45.07, 47.14	39.41, 43.93	2.0, 7.1
Missing, *n* (%)	12 (80.0)	3 (20.0)	30 (85.7)	5 (14.3)	42 (84.0)	8 (16.0)	
SF-12 Physical score							
N (%)	149 (81.0)	35 (19.0)	309 (84.9)	55 (15.1)	458 (83.6)	90 (16.4)	
Mean	38.2	36.7	37.1	37.6	37.5	37.2	0.3
95% CI	36.57, 39.90	33.72, 39.64	36.02, 38.16	35.25, 39.95	36.56, 38.37	35.41, 39.08	−1.9, 2.5
Missing, *n* (%)	12 (80.0)	3 (20.0)	30 (85.7)	5 (14.3)	42 (84.0)	8 (16.0)	
EQ-5D							
N (%)	156 (80.8)	37 (19.2)	329 (84.8)	59 (15.2)	485 (83.5)	96 (16.5)	
Mean	0.6	0.5	0.6	0.5	0.6	0.5	0.1*
95% CI	0.58, 0.66	0.39, 0.60	0.55, 0.61	0.41, 0.57	0.57, 0.62	0.42, 0.55	0.1, 0.2
Missing, *n* (%)	5 (83.3)	1 (16.7)	10 (90.9)	1 (9.1)	15 (88.2)	2 (11.8)	

Abbreviations: AM, active management; CBA, cognitive behavioural approach; CI, confidence interval; SF-12, Short Form 12.

*Indicates $0.001 \leq P$-value < 0.05; **indicates P-value < 0.001: P-value provided from the "data collection" main effect term in the linear regression model: response $= \beta_0 + \beta_1$ treatment $+ \beta_2$ data collection.

[a] Row percentages are based on the total respondents within each treatment arm and each row.

[b] Row percentages are based on the total respondents with each row.

compared with those who respond by telephone. There is a need to emphasize that these observed difference may well be affected by different types of biases, for example, participant selection and/or by the fact that two different methods of data collection were used. Use of the telephone as a supplementary method can introduce, for example, recall bias. Despite these disparities, the effect of the treatment difference over both data collection methods remained similar and therefore we can assume that there was nondifferential bias according to the treatment effect for the assessments.

This study has the following implications:

1. Maximization of response

The use of the telephone interviews as a supplementary method has proven to be an effective means of increasing response in this clinical trial. Using this method, the overall response rate increased by approximately 14%, making a substantial impact on the number of participants used in reporting the overall results. This brought the response rate up to 84% at 12 months follow-up for the BeST trial, which is impressive in comparison with follow-up in similar studies.

2. Increased generalizability and precision

The main trial results for BeST have been reported in Ref. [11], and the present study adds further strength to these results as it has been demonstrated that the overall treatment effects for the outcomes reported had not been affected by the method of data collection. The additional telephone follow up increased the generalizability and also meant that that the treatment effects were more precise compared with those obtained if only the postal questionnaire data had been used.

3. Completeness of data

There were fewer missing data reported for telephone respondents compared with those who responded by post and this was in accordance with other studies [22–24]. The high response rate in the telephone interview may have been because of the nature of this method and also because of the length of questionnaire, which was shorter and therefore quicker to complete compared with the postal questionnaire.

4. Strengths and limitations

In this article, there is evidence to suggest that researchers should consider different methods to collect follow-up data

Table 3. Number (and %) of questions missing and not missing for each outcome at 12 months

	AM		AM + CBA			Total	
Outcomes	**Postal**	**Telephone**	**Postal**	**Telephone**	**Odds ratio (postal vs. telephone)**	**Postal**	**Telephone**
Modified Von Korff Disability							
Missing, *n* (%)	15 (3.1)	1 (0.9)	30 (2.9)	0	—	45 (3.0)	1 (0.3)
Not missing, *n* (%)	468 (96.9)	113 (99.1)	987 (97.1)	180 (100.0)		1455 (97.0)	293 (99.7)
Modified Von Korff pain							
Missing, *n* (%)	10 (2.1)	0	9 (0.9)	0	—	19 (1.3)	0
Not missing, *n* (%)	473 (97.9)	114 (100.0)	1008 (99.1)	180 (100.0)		1481 (98.7)	294 (100.0)
SF-12 Physical							
Missing, *n* (%)	32 (1.7)	6 (1.3)	69 (1.7)	1 (0.1)	9.6 (1.09, 458.94)*	101 (1.7)	7 (0.6)
Not missing, *n* (%)	1900 (98.3)	450 (98.7)	3999 (98.3)	719 (99.9)		5899 (98.3)	1169 (99.4)
SF-12 Mental							
Missing, *n* (%)	32 (1.7)	6 (1.3)	69 (1.7)	1 (0.1)	9.6 (1.09, 458.94)*	101 (1.7)	7 (0.6)
Not missing, *n* (%)	1900 (98.3)	450 (98.7)	3999 (98.3)	719 (99.9)		5899 (98.3)	1169 (99.4)
EQ-5D							
Missing, *n* (%)	5 (0.6)	6 (3.2)	26 (1.5)	1 (0.3)	21.2 (1.96, >999.99)*	31 (1.2)	7 (1.4)
Not missing, *n* (%)	800 (99.4)	184 (96.8)	1669 (98.5)	299 (99.7)		2469 (98.8)	483 (98.6)

Abbreviations: AM, active management; CBA, cognitive behavioural approach; CI, confidence interval; SF-12, Short Form 12.

Note: Percentages are based on (number of missing questions/[number of questions × number of responders]) × 100.

*Indicates $0.001 \leq P$-value < 0.05; **indicates P-value < 0.001: P-value provided from the interaction term in the logistic regression model: $\ln(p/[1\text{-}p]) = \beta_0 + \beta_1$ treatment $+ \beta_2$ data collection $+ \beta_3$ data collection × treatment; where p = probability of missing item.

to maximize response rates. However, it is important to bear in mind that the use of different data collection methods may result in conclusions, which are biased as a result of factors introduced by this study, for example, patient selection and lack of similarities in the two methods used. However, provided that this bias is nondifferential over the treatment arms, as was the case in our study, then an assessment of treatment difference is possible. To test the effect of using different methods of data collection, a reliability study would be the ideal design. Each participant would be exposed to both methods and agreement in their response would be tested using, for example, the kappa statistics/correlation coefficient. However, in this setting, this was not possible as collecting data by telephone was used as a secondary method of data collection, to enhance the response rates. Also, there are limitations to collecting data by telephone in particular only certain information could be collected. By the very nature of this method, it was difficult to collect information on lengthy instruments or those that required the presence of the participant (e.g., the visual analog scale that required a tick mark to indicate the health status on a scale of 1–100). The total additional cost and time of making telephone calls when chasing participants for follow-up has not been documented in our study. We are therefore unable to quantify the time and cost.

5. Conclusion

We recommend the use of telephone follow-up to collect outcome data from nonresponders in clinical trials. This study has illustrated the importance of the use of two different methods of data collection in reducing loss to follow-up and thus retaining the power of the study. This in turn has yielded data of comparable quality and enhanced precision of effect estimates. One of the most important criteria for a successful clinical trial is retention of patients and their data. Many clinical studies fail because of lack of response provided by patients. We recommend that researchers consider the potential use of more than one method of data collection at the initial stages of funding applications. It is essential that the supplementary method be stated in the protocol and thus funded adequately. However, if different

Table 4. Internal consistency for the outcome at 12 months using Chronbach's alpha

	AM		AM + CBA		Total	
Outcomes	**Postal**	**Telephone**	**Postal**	**Telephone**	**Postal**	**Telephone**
Modified Von Korff Disability	0.92	0.87	0.94	0.90	0.93	0.89
Modified Von Korff Pain	0.93	0.89	0.94	0.91	0.94	0.91
SF-12 Physical	0.92	0.88	0.93	0.92	0.93	0.91
SF-12 Mental	0.92	0.88	0.93	0.92	0.93	0.91
EQ-5D	0.72	0.57	0.75	0.69	0.74	0.64

Abbreviations: AM, active management; CBA, cognitive behavioural approach; CI, confidence interval; SF-12, Short Form 12.

Table 5. Outcome measures summarized using change from baseline to 12 months

	AM[a]		AM + CBA[a]		Total[b]		
Core outcomes	**Post**	**Telephone**	**Post**	**Telephone**	**Postal**	**Telephone**	**Difference (postal–telephone)**
Modified Von Korff Disability							
N (%)	148 (80.0)	37 (20.0)	306 (83.8)	59 (16.2)	454 (82.5)	96 (17.5)	
Mean	7.7	0.09	15.7	8.08	13.0	5.4	7.6*
95% CI	4.22, 11.11	−5.16, 5.33	13.19, 18.13	3.15, 13.01	10.87, 15.07	0.73, 10.06	2.4, 12.7
Modified Von Korff Pain							
N (%)	155 (80.3)	38 (19.7)	329 (84.8)	59 (15.2)	484 (83.3)	97 (16.7)	
Mean	8.2	3.3	15.0	10.1	12.7	7.8	4.9
95% CI	4.99, 11.40	−1.62, 8.28	12.69, 17.23	5.42, 14.76	10.78, 14.65	3.42, 12.27	−0.01, 9.8
SF-12 Physical score							
N (%)	138 (80.2)	34 (19.8)	293 (84.2)	55 (15.8)	431 (82.9)	89 (17.1)	
Mean	−1.4	0.2	−5.3	−3.7	−4.0	−2.4	−1.6
95% CI	−2.69, −0.10	−1.74, 2.20	−6.22, −4.39	−5.52, −1.83	−4.79, −3.23	−4.14, −0.63	−3.6, 0.3
SF-12 Mental score							
N (%)	138 (80.2)	34 (19.8)	293 (84.2)	55 (15.8)	431 (82.9)	89 (17.1)	0.802
Mean	−0.9	−0.6	−1.1	−0.8	−1.0	−0.7	−0.3
95% CI	−2.38, 0.57	−2.86, 1.62	−2.10, −0.02	−2.89, 1.34	−1.89, −0.12	−2.73, 1.28	−2.5, 1.9
EQ-5D							
N (%)	154 (81.5)	35 (18.5)	322 (85.0)	57 (15.0)	476 (83.8)	92 (16.2)	
Mean	0.01	−0.06	−0.05	−0.12	−0.03	−0.10	0.1*
95% CI	−0.03, 0.05	−0.12, −0.01	−0.08, −0.02	−0.18, −0.06	−0.05, −0.01	−0.15, −0.05	0.01, 0.1

Abbreviations: AM, active management; CBA, cognitive behavioural approach; CI, confidence interval; SF-12, Short Form 12.

*Indicates $0.001 \leq P$-value < 0.05; **indicates P-value < 0.001: P-value provided from the "data collection" main effect term in the linear regression model: response $= \beta_0 + \beta_1$ baseline $+ \beta_2$ age $+ \beta_3$ sex $+ \beta_4$ treatment $+ \beta_5$ data collection.

[a] Row percentages are based on the total respondents within each treatment arm and each row.

[b] Row percentages are based on the total respondents with each row.

methods are used to collect data in future trials, then the data need to be carefully scrutinized and bias needs to be carefully assessed.

References

[1] Siemiatycki JA. Comparison of mail, telephone, and home interview strategies for household health surveys. Am J Public Health 1979;69:238–45.

[2] Siemiatycki J, Campbell S. Nonresponse bias and early versus all responders in mail and telephone surveys. Am J Epidemiol 1984;120:291–301.

[3] Cook K, Choi S, Crane P, Deyo R, Johnson K, Amtmann D. Letting the CAT out of the Bag: Comparing Computer Adaptive Tests and an 11-item Short form of the Roland Morris Disability Questionnaire. Spine 2008;33:1378–83.

[4] Bowling A. Model of questionnaire administration can have serious affects on data quality. J Public Health 2005;27:281–91.

[5] Edwards P, Roberts I, Clarke M, DiGuiseppi C, Wentz R, Kwan I, et al. Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst Rev 2009;(3). MR000008.

[6] Nakash R, Hutton J, Jorstad-Stein E, Gates S, Lamb S. Maximising response to postal questionnaires- A systematic review of randomised trials in health research. BMC Med Res Methodol 2006;6:5.

[7] Ivens M, Johansen C, Skov T. Comparison of a self-administered questionnaire and a telephone interview of 146 Danish waste collectors. Am J Ind Med 1997;31(5):653–8.

[8] Brambilla D, McKinlay S. A comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey. Am J Epidemiol 1987;126:962–71.

[9] Sibbald B, Addington-Hall J, Brennenman D, Freeling P. Telephone versus postal surveys of general practitioners: methodological considerations. Br J Gen Pract 1994;44:297–300.

[10] Parker C, Dewey M. Assessing research outcomes by postal questionnaires with telephone follow-up. Int J Epidemiol 2000;29:1065–9.

[11] Lamb SE, Hansen Z, Lall R, Castelnuovo E, Withers EJ, Nichols V, et al; Back Skills Training Trial investigators. Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. Lancet 2010;375:916–23.

[12] Lamb SE, Lall R, Hansen Z, Castelnuovo E, Withers EJ, Nichols V, et al., on behalf of the BeST trial group. A multi-center randomized controlled trial of a primary care based cognitive behavioral program for low back pain (the Back Skills Training Trial): HTA Report 01/75/01. Commissioned Monograph (ISRCTN37807450); 2009.

[13] Roland M, Waddell G, Klaber Moffett J, Burton AK, Main CJ, Cantrell T. The back book. Norwich, UK: Stationary Office; 2005.

[14] Hansen Z, Daykin A, Lamb SE. A cognitive-behavioural programme for the management of low back pain in primary care: a description and justification of the intervention used in the Back Skills Training Trial (BeST; ISRCTN 54717854). Physiotherapy 2010;96:87–94.

[15] Von Korff M, Ormel J, Keefe F, Dworkin S. Grading the severity of chronic pain. Pain 1992;50:133–49.

[16] Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med care 1996;34:220–33.

[17] EuroQol Group. EuroQol-A new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.

[18] Ohio University. MicroSoft ACCESS Reference guide 2003. Ohio University Computer Services Centre; 2006.

[19] Sheh LZ. A practical and efficient approach to Database Quality Audit in Clinical Trials. Drug Inf J 2006;40(4):385–93.

[20] StataCorp. Statistical software: release 7.0. College Station, TX: Stata Corporation; 2001.
[21] CONSORT statement. Available at http://www.consort-statement.org/consort-statement/flow-diagram0/.
[22] de Leeuw ED. Mode effects in survey research: a comparison of mail, telephone and face to face surveys. Bull Methodol Sociol 1993;43:3–19.
[23] Perkins J, Sanson-Fisher R. An examination of self- and telephone-administered models of administration for the Australian SF-36. J Clin Epidemiol 1998;51:969–73.
[24] Feveile H, Olsen O, Hogh A. A randomised trial of mailed questionnaire versus interviews: Response patterns in a survey. BMC Med Res Methodol 2007;7:27. doi: 10.1186/1471-2288-7-27.