Original Article

Performance status and trial site-level factors are associated with missing data in palliative care trials: An individual participant-level data analysis of 10 phase 3 trials

Jamilla A Hussain[1] iD **, Ian R White[2], Miriam J Johnson[3]** iD **, Martin Bland[4] and David C Currow[5]** iD

Abstract

Background: Missing data compromise the internal and external validity of trial findings, however there is limited evidence on how best to reduce missing data in palliative care trials.

Aim: To assess the association between participant and site level factors and missing data in palliative care trials.

Design and setting: Individual participant-level data analysis of 10 phase 3 palliative care trials using multi-level cross-classified models.

Results: Participants with missing data at the previous time-point and poorer performance status were more likely to have missing data for the primary outcome and quality of life outcomes, at the primary follow-up point and end of follow-up. At the end of follow-up, the number of site randomisations and number of study site personnel were significantly associated with missing data. Trial duration and the number of research personnel explained most of the variance at the trial and site-level respectively, except for the primary outcome where the amount of data requested was most important at the trial-level. Variance at the trial level was more substantial than at the site level across models and considerable variance remained unexplained for all models except quality of life at the end of follow-up.

Conclusion: Participants with a poorer performance status are at higher risk of missing data in palliative care trials and require additional support to provide complete data. Performance status is a potential auxiliary variable for missing data imputation models. Reducing trial variability should be prioritised and further factors need to be identified and explored to explain the residual variance.

Keywords

Missing data, randomised controlled trials, palliative care, palliative medicine, quality of life, research personnel, lost to follow-up

---

What is already known

- Missing data reduce the power, precision, generalisability and validity of study findings.
- A systematic review of palliative care trials found nearly one quarter of primary outcome data are missing.
- It is essential that missing data are reduced as much as possible but how to effectively achieve this is unknown.

What this paper adds

- Using individual participant-level data, this study found a poorer performance status and missing data at a previous time-point are strongly and consistently associated with missing data in palliative care trials.
- Site-level factors were also found to have a significant association with missing data at the end of follow-up, although variability between trials was more substantial than between sites.
- Trial duration and the number of research personnel explained most of the variance at the trial and site-level respectively, except for the primary outcome at the primary follow-up point where the amount of data requested was most important at the trial-level.

---

[1]Wolfson Palliative Care Research Centre, Hull York Medical School, Hull, UK
[2]MRC Clinical Trials Unit at UCL, London, UK
[3]Wolfson Palliative Care Research Centre, University of Hull, Hull, UK
[4]Health Sciences Department, University of York, York, UK
[5]IMPACCT, Faculty of Health, University of Technology Sydney, Ultimo, NSW, Australia

Corresponding author:
Jamilla A Hussain, Wolfson Palliative Care Research Centre, Allam Medical Building, University of Hull, Cottingham Road, Hull HU6 7RX, UK.
Email: hyjah@hyms.ac.uk

Implications for practice, theory or policy

- Participants with a poorer performance status and those with previous missing data are at higher risk of missing data in palliative care trials and should be identified early and provided with additional support to enable the provision of complete data; they should not be excluded from trials.
- Performance status, in particular, could also be considered as an auxiliary variable for missing data imputation models in palliative care trials thus making a missing at random assumption more plausible and missing data analyses more robust.
- Reducing variability between trials is important and further assessment of how site-level factors affect missing data is required.

Introduction

Missing data compromise the power, precision, generalisability and validity of study findings. A systematic review of palliative care trials found nearly one quarter of primary outcome data were missing with evidence of subsequent bias.[1]

To minimise the impact of missing data on trial findings, prevention is important as statistical methods to handle existing missing data are based on unverifiable assumptions.[2,3] However, there is little research on how best to reduce missing data.[2] A Cochrane review of trials testing strategies to improve retention in randomised controlled trials, an overlapping issue, included 38 studies that assessed a range of interventions.[4] Assessment of interventions to improve trial management and site-level factors was limited, and most were evaluated in single trials in a particular context.[4]

Effective interventions to reduce missing data should be based on evidence. It is necessary to identify the factors associated with missingness to inform the design of such interventions. A meta-regression of factors associated with missing data in a systematic review of palliative care trials found that trial duration and the amount of data requested during the trial were associated with missing data.[1] However, there was insufficient evidence that participant-level factors such as age and performance status were associated with missing data.[1] This analysis, however, relied on aggregate-level participant data. Furthermore, there were no data regarding site-level factors.

Individual-participant level data on the other hand uses the raw unit-level data from each primary study.[23,24] This allows different sources of heterogeneity in the effect estimate to be explored (e.g. participant, site and trial-level), multiple participant-level factors to be examined in combination, identification and handling of missing data at the individual-level and models to be developed and validated using statistical techniques that are standardised across studies.[5]

The aim of this study was to use individual participant-level data to assess participant and site-level factors associated with missing data in palliative care trials. The objectives were to:

1. Assess factors associated with missing data for the primary outcome at the primary follow-up point (Timepoint 1).
2. Assess factors associated with missing data for any primary quality of life (QoL) outcome at Timepoint 1, given its importance to palliative care clinical practice.[6]
3. Assess factors associated with missing data for the primary outcome and QoL outcome at the end of follow-up (Timepoint 2).

Methods

Protocol

The protocol for this review could not be registered with PROSPERO,[7] as it was a methodological review and did not meet the requirements for registration. The protocol was internally peer-reviewed by methodological experts including those with expertise in individual participant level data analysis.

Eligibility criteria

Randomised controlled trials were eligible if: participants were over 18 years old with an advanced life-limiting illness and palliative care needs; the interventions were palliative where the primary aim is to improve QoL, rather than survival, although this may be a secondary gain; the comparator was a placebo, standard care or another palliative intervention; the primary outcome was patient-reported; the trial was a priori adequately powered and the trial was completed in the 5 years before this study began. Published and unpublished trials were included with no language restrictions.

Identifying studies, selection and data collection process

Trials were identified through professional contacts. The level of missing data was unknown to the principal investigator (JH) before contact and did not influence whether the trial protocol was assessed. All 10 anonymous datasets were securely accessed.

Table 1. Explanatory variables assessed.

Level	Explanatory variable
Participant	• Missingness for the primary outcome or quality of life outcome at the previous time-point • Australia-Modified Karnofsky Performance Scale (AKPS) at the previous time-point (i.e. at Timepoint 0 = T0-AKPS, at Timepoint 1 = T1-AKPS) • Age • Diagnosis (dichotomised to cancer and non-cancer).
Site	Based on in-depth interviews with research staff and administrators involved in palliative care trials and the recommendations on how to minimise missing data by the National Research Council[2]: • Site randomisations: number of participants randomised by the site for the 10 trials included in the IPD analysis • Research personnel: number of research personnel working at the site on palliative care trials, across the course of the trials • Site coordinator: whether the site had a coordinator (1 = yes, 0 = no) • Site home visits: whether the site could conduct home visits (1 = yes, 0 = no) • Site experience: the level of palliative care trial experience of the site as judged by the trial coordinator and/or chief investigator (1 = little experience, 2 = moderate experience, 3 = very experienced).
Trial	A systematic review of palliative care trials found that trial duration and items of data requested were associated with missing data.[1] The Individual participant-level data analysis aimed to determine whether participant and site-level factors were associated with missing data once these trial-level factors were taken into account: • Trial duration at Timepoint 1 and Timepoint 2 • Amount of data requested from participants at Timepoint 1 and Timepoint 2.

Data items

The participant, site and trial-level data items extracted are presented in Table 1. The primary outcome and QoL data were extracted at Timepoint 0, Timepoint 1 and Timepoint 2 – defined as baseline (Timepoint 0), primary follow-up point (Timepoint 1) and end of follow-up (Timepoint 2). If the QoL measure was not the primary outcome, the data for the main QoL measure of interest were extracted. The Australia-Modified Karnofsky Performance Scale (AKPS) was extracted at Timepoint 0 and Timepoint 1.

Specification of outcomes and explanatory variables

The pre-specified outcome variables of interest were whether the primary outcome and QoL scores were observed or missing at Timepoint 1 and Timepoint 2:

- Timepoint1-PO-missing = whether the primary outcome was missing at Timepoint 1
- Timepoint2-PO-missing = whether the primary outcome was missing at Timepoint 2
- Timepoint1-QoL-missing = whether the QoL outcome was missing at Timepoint 1
- Timepoint2-QoL-missing = whether the QoL outcome was missing at Timepoint 2.

For the primary outcome, a single symptom-control measure was used for all the trials (this was not a pre-specified criteria) therefore whether this value was entered in the dataset or not was coded as 1 = missing and 0 = observed.

The QoL measures consisted of a number of question items (range 20–28 items), which were found, in general, to be either all answered or all missing. Therefore, these outcomes were dichotomised into all present or absent if one or more questions were missing. Specifications of the explanatory variables are available in Table 1. A systematic review of palliative care trials found that trial duration and the individual number of questions and tests requested from the participant (as a measure of trial burden) were associated with missing data.[1] The models adjusted for these two variables as the individual participant-level data analysis aimed to determine whether participant and site-level factors were associated with missing data once these trial-level factors were taken into account.

Synthesis methods

Multilevel, cross-classified models were developed as participants (level 1 units) were nested within combinations of trials and sites (level 2 units). A mixed-effects model was used, with fixed-effects for all independent variables and random intercepts for the trials and sites.[8]

Analysis strategy. The analysis strategy was based on a systematic approach that commenced with level 1 fixed-effects, then the addition of higher-level explanatory variables, followed by tests for interactions.[9,10] Further details of the analysis strategy are available in Supplemental Material 1.

Categorising variables. To determine whether continuous explanatory variables should be treated as categorical variables, the relationship between the explanatory

variable and outcome was assessed using a scatter plot. If this indicated that categorisation of the variable might fit the data better, the model treating the variable as a continuous variable was compared to that treating it as a categorical variable using a likelihood ratio test.

Variance. The proportion of the total variance due to the different group-levels was assessed using the variance partition coefficient. The variance partition coefficient is interpreted as the proportion of the total residual variance in the propensity to be missing that is due to differences between either trials or sites, or both.[11] In this analysis the latent variable representation approach was used.[12]

In each model there was evidence of many combinations of trials and sites with several observations, therefore a random interaction between trial and site was also tested using a likelihood ratio test[13] and, if appropriate, included as a further level. This allowed the assumption that the trial and site effects were additive to be relaxed.[13]

Interactions. Interactions of both age and diagnosis with AKPS were tested.

Handling missing data in explanatory variables. There were no missing data for the model outcomes as missingness was the outcome of interest. Missing values for the explanatory variables were explored to determine the justifiability of a missing completely at random assumption. If this was not found, imputation using chained equation models congenial with the model of interest were conducted under missing at random and plausible missing not at random assumptions and data were imputed using within-trial imputation. The effect estimates and random-effects were compared as part of principled missing data sensitivity analyses (see Supplemental Material 2).

All analyses were conducted using Stata v.13 and a *p*-value ⩽0.05 was considered to be statistically significant unless otherwise stated. Data extraction and analyses were completed in December 2017.

Ethics and consent. Secondary analysis of anonymised data of the included trial datasets was allowed under the original human research ethics approval for each trial.

Results

Thirteen studies were screened. Ten were eligible for at least one model, all of which were conducted in Australia and the UK (see Table 2). One trial was excluded (feasibility study) and data were not provided for another two studies. The number of trials included for each model varied because of varying measurement outcomes and timepoints (see Table 3). However, the descriptive statistics of the variables included in the models were comparable (Table 3).

Factors associated with missing data at Timepoint 1

The multivariable models for missing data for the primary outcome at Timepoint 1 showed a strong association between baseline and Timepoint 1 primary outcome missing data (OR 17.2, 95% CI: 8.6, 34.5). This indicates that those with missing data at baseline were highly likely to have missing data at Timepoint 1 (Table 4). As the baseline AKPS increased (i.e. improved), the odds of missing data for the primary outcome at Timepoint 1 reduced significantly (OR 0.78 (95% CI: 0.70, 0.87) per 10-unit increase) (Table 4).

Findings for Timepoint1-QoL-missing were similar. As QoL was a secondary outcome in all trials (not a pre-specified criteria), the association with missing data for the primary outcome at baseline was also assessed, which found participants with missing data for the primary outcome at baseline were more likely to have missing data for the QoL outcome at Timepoint 1 (Table 4).

Factors associated with missing data at Timepoint 2

When the variable site-personnel was treated as a categorical variable (vs continuous), it fitted the data better (Timepoint2-PO-missing *p* = 0.0009; Timepoint2-QoL-missing *p* = 0.0002) and was therefore treated as categorical (Table 5, Supplemental Material 3).

A strong association was found between missing data for the primary outcome at Timepoint 2 and Timepoint 1 (OR 8.0, 95% CI: 5.4, 11.8) and Timepoint1-AKPS (per 10 unit change OR 0.7, 95% CI: 0.6, 0.8); which indicates that those with previous missing data for the primary outcome and poorer performance status were more likely to have missing data at Timepoint 2. Sites that randomised more participants were more likely to have missing data (per 10 randomisations OR 1.1, 95% CI: 1.0, 1.2). The number of site personnel was also significantly associated with missing data, with sites who had two research personnel being more likely to have missing data (OR 2.6, 95% CI: 1.1, 6.0), and those with four personnel less likely (OR 0.07, 95% CI: 0.01, 08) to have missing data, than sites with one research nurse.

Findings for missing QoL data at Timepoint 2 were similar, with an additional strong association with missing data for the primary outcome at Timepoint 1 (OR 11.8, 95% CI: 6.9, 20.3) and trial duration (per 7 days, OR 0.6, 95% CI: 0.5, 0.8).

For all models, there was insufficient evidence of significant interactions between participant-level factors (data not shown).

Table 2. Characteristics of included trials.

Trial	1	2	3	4	5	6	7	8	9	10
Country										
Australia	✓	✓	✓	✓	✓	✓			✓	✓
United Kingdom							✓	✓	✓	
Published as of 02/06/2016	✓	X	✓	X	X	✓	✓	✓	X	X
Trial design										
Parallel	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Cross-over										✓
No. of trial arms	2	3	2	3	2	2	2	2	2	2
Number of sites	9	10	12	12	14	8	8	1	8	3
Intervention										
Pharmacological	✓	✓	✓	✓	✓	✓			✓	✓
Complex							✓	✓		
Control										
Placebo	✓	✓	✓	✓	✓	✓				
Training							✓			
Standard care								✓	✓	
Pharmacological										✓
Primary outcome										
Symptom	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
QoL										
Trial duration										
Timepoint 1 (days)	6	3	4	7	8	7	28	42	22	28
Timepoint 2 (days)	33	31	32	35	36	NA	56	84	50	NA
Participants										
No. randomised	185	247	112	176	354	257	156	154	101	104
Age (years)*	63.6	75.0	65.2	72.5	71.6	73.5	69.2	70.7	67.0	62.7
Male (%)	56.6	65.4	16.7	60.6	63.3	63.1	60.1	51.6	60.0	57.7
Diagnosis (%):										
Cancer	100	94.5	100	100	41.1	14.9	100	44.2	94.0	100
Respiratory		0.8			49.7	77.8		55.2	3.0	
Cardiovascular		3.4			9.2	5.0		0	2.0	
Other		1.3				2.3		0.7	1.0	
T0-AKPS*	56.3	41.4	51.6	62.9	61.6	65.4	70.3	69.0	62.4	71.9

*Baseline Australia-modified Karnofsky Performance Scale.

Variance explained

A non-additive model with a trial-by-site interaction level was the preferred model for missing data for the primary outcome at Timepoint 1 ($p = 0.005$) and Timepoint 2 ($p = 0.01$) (Table 6). By adding the interaction level to the Timepoint1-PO-missing null model, the variance at the site-level became negligible (<0.0001 on the log-odds scale). This suggests that for the primary outcome at Timepoint 1 the effect of sites differs within trials, rather than sites having an independent effect on missingness regardless of the trial for which they were recruiting.

The multivariable model explained almost all of the variance at the trial and site-level for the Timepoint2-QoL-missing model but not for the other outcomes (Table 6). For the Timepoint1-PO-missing model, data requested explained the trial-level variability the most (Supplemental Material 4). Trial duration explained the trial-level, and the number of research personnel the site-level, variance the most for Timepoint2-PO-missing, Timepoint1-QoL-missing and Timepoint2-QoL-missing models (Supplemental Material 4).

Explanatory variable missing data

Exploration of the missing data suggested complete case analysis under a MCAR assumption was reasonable for the Timepoint1-PO-missing model. However, the MAR assumption was more plausible for Timepoint1-QoL-missing, Timepoint2-PO-missing and Timepoint2-QoL-missing, therefore the final model for these outcomes used multiple imputation under a MAR assumption. Missing data sensitivity analyses under various MNAR assumptions did not change the findings considerably (Supplemental Material 2).

Table 3. Summary of the variables included in each model.

Explanatory and outcome variables		Model			
		Timepoint1-PO-missing	Timepoint2-PO-missing	Timepoint1-QOL-missing	Timepoint2-QOL-missing
		Percentage/mean (SD, range)	Percentage/mean (SD, range)	Percentage/mean (SD, range)	Percentage/mean (SD, range)
Participant-level	T0-PO-missing	6.6%	4.4%	5.8%	4.0%
	T1-PO-missing	25.8%*	25.9%	–	26.6%
	T2-PO-missing	–	54.0%*	–	–
	T0-QOL-missing	–	–	33.4%	25.5%
	T1-QOL-missing	–	–	39.0%*	29.6%
	T2-QOL-missing	–	–	–	54.7%*
	Age	70.2 (11.3, range 20, 97)	69.2 (11.2, range: 20, 94)	69.9 (11.1, range 20, 94)	69.6 (10.9, range 20, 94)
	T0-AKPS	60.4 (15.0, range 20, 100)	62.2 (13.10, range 20,100)	62.8 (12.8, range 20, 100)	63.2 (12.7, range 20, 100)
	T1-AKPS	–	57.3 (16.6, range 0, 100)	–	61.6 (15.1, range 0, 100)
	Cancer diagnosis	70.9%	74.0%	65.0%	71.4%
Trial-level	T1-Duration (days)	15.1 (11.9, range 3, 42)	–	13.7 (12.0, range 4, 42)	–
	T1-Data requested	653.5 (253.7, range 317, 1049)	–	591.8 (226.2, range 317, 1049)	–
	T2-Duration (days)	–	44.7 (16.8, range 32, 84)	–	45.9 (17.1, range 33, 84)
	T2-Data requested	–	1224.7 (434.2, range 539, 1872)	–	1190.4 (440.2, range 539, 1872)
Site-level	Site randomisations	167.6 (137.7, range 1, 408)	152.6 (126.1, range 2, 408)	162.5 (136.5, range 1, 408)	148.8 (125.9, range 2, 408)
	No. site personnel				
	1	20.1%	23.6%	24.3%	25.7%
	2	31.0%	30.3%	30.3%	30.6%
	3	26.7%	29.3%	24.3%	27.4%
	4	22.2%	16.8%	21.1%	16.2%
	Site coordinator	76.2%	69.2%	73.5%	67.0%
	Site home visit possible	68.7%	75.3%	70.5%	77.4%
	Site experience				
	1 (low-level)	22.4%	17.7%	21.8%	17.5%
	2 (moderate-level)	22.9%	30.3%	25.3%	32.7%
	3 (high-level)	54.8%	52.0%	52.9%	49.8%
Number of trials included		10	7	8	6

PO: primary outcome; QoL: quality of life; T0: baseline; T1: timepoint 1; T2: timepoint 2; SD: standard deviation.
*Outcome variable for the model.

Discussion

Individual participant-level data analysis of 10 palliative care trials found participants with a poorer performance status and those with previous missing data were more likely to have missing data at both Timepoint 1 and Timepoint 2. At Timepoint 2, site level factors were also found to be significantly associated with missing data.

Trial duration and the number of research personnel explained most of the variance at the trial and site-level respectively, except for the primary outcome at Timepoint 1 where the amount of data requested was most important at the trial-level. Variance at the trial level was more substantial than at the site level and a considerable proportion of the variance remained unexplained at the trial and site-level for most models.

Table 4. Multivariable multi-level model: participant, trial and site level factors associated with missing data for the primary outcome and main QoL outcome at Timepoint1.

Fixed effects	Timepoint1-PO-missing		Timepoint1-QoL-missing	
	Odds ratio	95% CI	Odds ratio	95% CI
Timepoint0-QoL-missing	–	–	2.86***	2.06–3.96
Timepoint0-PO-missing	17.19***	8.55–34.53	2.56*	1.11–5.90
AKPS at Timepoint0 (per 10-units)	0.78***	0.70–0.87	0.78***	0.70–0.89
Cancer diagnosis	1.44	0.94–2.20	1.25	0.80–1.94
Age (per 10-years)	0.99	0.88–1.12	1.01	0.89–1.13
Trial duration to Timepoint1 (per 7 days)	1.19	0.88–1.61	1.03	0.67–1.60
Trial data requested to Timepoint1 (per 30 items)	1.03	0.97–1.09	0.99	0.90–1.08
Site randomisations (per 10)	0.98	0.96–1.01	1.00	0.96–1.04
No. of site personnel	1.49	0.93–2.38	1.65	0.97–2.81
Site coordinator	0.82	0.41–1.65	0.89	0.35–2.29
Site home visit	0.81	0.50–1.33	0.64	0.32–1.28
Site experience	0.85	0.63–1.15	0.77	0.52–1.14

PO: primary outcome; QoL: quality of life; Timepoint0: baseline; Timepoint1: primary end-point.
*$p < 0.05$. ***$p < 0.001$.

Table 5. Multivariable multi-level model: participant, trial and site level factors associated with missing data for the primary outcome and main QoL outcome at Timepoint 2.

Fixed effect	Timepoint2-PO-missing		Timepoint2-QoL-missing	
	Odds ratio	95% CI	Odds ratio	95% CI
Timepoint1-PO/QoL-missing	7.95***	5.36–11.81	2.16***	1.39–3.36
Timepoint1-PO-missing	–	–	11.79***	6.86–20.26
AKPS at Timepoint1 (per 10-units)	0.73***	0.64–0.82	0.79***	0.69–0.90
Cancer diagnosis	1.49	0.96–2.32	1.40	0.91–2.16
Age (per 10-years)	0.94	0.83–1.08	1.02	0.88–1.18
Trial duration to Timepoint2 (per 7 days)	0.81	0.52–1.26	0.59***	0.45–0.78
Trial data requested to Timepoint2 (per 30 questions)	0.97	0.90–1.03	0.98	0.94–1.02
Site randomisations (per 10)	1.08*	1.01–1.16	1.15***	1.09–1.21
No. of site personnel	$p = 0.01$*		$p < 0.0001$***	
1	1 (reference)		1 (reference)	
2	2.59*	1.11–6.02	1.88	0.97–3.63
3	1.48	0.51–4.30	0.96	0.40–2.31
4	0.07*	0.01–0.84	0.01***	<0.01–0.08
Site coordinator	1.95	0.71–5.31	0.68	0.28–1.64
Site home visit	0.54	0.27–1.07	1.60	0.89–2.90
Site experience	1.39	0.95–2.03	1.22	0.90–1.66

PO: primary outcome; QoL: quality of life; Timepoint1: primary end-point; Timepoint2: end of follow-up.
*$p < 0.05$. ***$p < 0.001$.

Participant-level factors

Missingness of the primary outcome and principal QoL measure at the previous time-point was strongly and consistently associated with missingness at the time-point of interest. This is most likely driven by complete withdrawals at the previous time-point. However, 17.3% of participants with missing data at a previous time-point provided data at the following time-point of interest. Thus, for participants continuing in a trial, missing data at one time-point should be a 'red flag' for future missing data.

Participants who were more functionally limited were more likely to have missing data. Trialists should not use this to justify the exclusion of participants with poor performance status from palliative care trials in order to reduce missing data. Such patients are a core group who require palliative care input. If eligible for the intervention in clinical practice, it is essential that they are actively

Table 6. Multivariable multi-level model: residual variance, variance partition coefficient (VPC) and proportion of variance explained at the different levels.

Level		Timpoint1-PO-missing	Timepoint2-PO-missing	Timepoint1-QoL-missing	Timepoint2-QoL-missing
Trial	Final model: Variance (95% CI); VPC[a]%	0.44 (0.15, 1.28); 11.0%	0.41 (0.11, 1.64); 10.4%	0.91 (0.32, 2.63); 20.3%	0.12 (0.02, 0.81); 3.5%
	Null model: Variance (95% CI); VPC[a]%	0.55 (0.19, 1.56); 13.6%	0.57 (0.12, 2.58); 12.7%	1.03 (0.36, 2.96); 21.6%	1.03 (0.28, 3.73); 21.4%
	Proportion of variance explained[b]	18.5%	28.1%	11.7%	88.4%
Site	Final model: Variance (95% CI); VPC[a]%	0	0	0.29 (0.11, 0.75); 6.5%	0
	Null model: Variance (95% CI); VPC[a]%	0	0.32 (0.08, 1.31);	0.44 (0.21, 0.95); 9.3%	0.49 (0.21, 1.18); 10.3%
	Proportion of variance explained	–	100%	34.1%	100%
Trial-by-site interaction	Final model: Variance (95% CI); VPC[a]%	0.26 (0.10, 0.66); 6.4%	0.29 (0.12, 0.73); 7.3%		
	Null model: Variance (95% CI); VPC[a]%	0.35 (0.16, 0.75); 4.0%	0.31 (0.09, 1.08); 6.9%		
	Proportion of variance explained	25.7%	**6.5%**		

[a]Variance partition coefficient (VPC): Proportion of the total variance due to the different group levels that is trial and site.
[b]Proportion of the variance explained by the multivariable model compared to the null model (i.e. without covariates).

recruited to trials and supported to provide data to maximise the generalisability of trial findings. This study highlights the need for additional consideration on how best to support this group to provide data as a trial proceeds, this may include a more flexible study design, different modalities of data collection and the use of proxies.[14] Any interventions to reduce missing data need to be evaluated to determine the most effective and cost-effective measures through studies across trials.[14,15]

This participant-level data analysis is the first to systematically assess the impact of the AKPS on missing data and to demonstrate a consistent and robust association. AKPS is therefore potentially a useful auxiliary variable[16,17] for use in missing data imputation models in palliative care trials to make a missing at random assumption more plausible. Previous studies using aggregate level data[1] and participant-level data analysis[18] have not demonstrated an association between performance status and missing data. However this is likely related to ecological bias[1] and use of less sensitive measures[18] in these studies.

Site-level factors

Site-level variables and missing data were significantly associated at Timepoint 2 but not Timepoint 1. This may be because intensive central monitoring and checks are often in place for outcomes at Timepoint 1, but not always so at Timepoint 2 due to limited resources. The impact of site-level practices therefore may be more evident and influential after Timepoint 1 as the burden of the study on participants and site-staff increase. At Timepoint 2, sites that recruited more participants were more likely to have

missing data. These findings are new, and the underlying reasons for these need exploration.

The number of research personnel employed at a site and missingness at Timepoint 2 were significantly associated. This was not a dose-response or simple linear relationship, suggesting that there may be other influential site-level factors, which have not been included in the models; for example, whether the researchers work full/part-time, staff turnover, research experience and level and content of research training. Furthermore, there is recognition that conducting palliative care research can be emotionally challenging with a need for additional resources to promote job satisfaction and sustainability for trial staff[19] which may also play an important role in data quality. Further research into how the number of researchers and research culture at a site may influence missing data is required.

Residual variance

There was significant variance between trials and sites, indicating that the effect of trial and site factors were important and need to be addressed. This is an important finding as often the dominant focus of missing data literature in palliative care research is on participant-level factors such as participants' poor health and fatigue.[20] Also, for missing data for the primary outcome at Timepoint 1, unlike the other outcomes, there was little evidence that some sites were worse than others in a consistent way across trials (site-level variance), suggesting efforts to reduce missing data for the primary outcome at Timepoint 1 should focus on reducing between-trial variability.

The findings suggest that duration and the numbers of research personnel are key factors for consideration when trying to reduce missing data in palliative care trials, however other factors, such as data-requested, may be more important for the primary outcome at Timepoint 1. The variables included in the models explained almost all the variance at the trial and site-level for Timepoint2-QoL-missing, suggesting these variables have the greatest impact for QoL outcome missing data at Timepoint 2. However, as the Timepoint 2 models estimated a greater number of parameters at the site-level, this may be due to over-saturation at the site-level. Other participant, trial and site level factors, that were not included in the models, may be more crucial for the other outcomes and need to be investigated.

Limitations and strengths

The included trials were a small convenience sample, manageable within the time-frame of the study, which may limit the generalisability of the findings. However, the average extent of missing data for the primary outcome at Timepoint 1 (26%) and participant characteristics are consistent with a systematic review of 108 palliative care trials.[1] The principal investigator (JH) was blind to the extent and risk of bias of missing data and study characteristics at the time of selection, and the included trials involved participants with a range of ages, diagnoses and performance status scores, and varied in duration thus optimising generalisability (Table 2). Although sought, data on ethnicity and socio-economic status at the participant-level were not available consistently across the trials limiting our understanding of the representativeness of the study sample and the effect of these constructs on missing data. Data for two trials were not made available, and if the reason for this is related to missing data, it could bias the findings.[21] The majority of trials included in the sample were pharmacological trials and although two non-pharmacological complex intervention trials were included, additional considerations may be required for trials involving several interacting components.[22] The variables used in the models were restricted to those that were collected consistently across trials and could be quantified reliably.

The strengths of this study include the rigorous methodological approach which included multi-level modelling. Both published and unpublished palliative care trials were included and the participant-level data allowed the integrity of the data to be assessed.

Conclusion

Participants with a poorer performance status and previous missing data are at higher risk for missing data in palliative care trials and require early identification and support to provide complete data. These factors could also be considered as auxiliary variables in missing data imputation models, especially if associated with the missingness outcome. However, performance status only explained part of the residual variance, indicating that other factors affect missing data at the trial level; identifying these factors may help to reduce missing data in future trials, especially if they are modifiable factors.

Author contribution

JH, DC, MJ, MB conceived the idea of the study. JH was the principal investigator and developed the protocol, conducted the data extraction and analysis and wrote the first draft of the paper. IW developed the data analysis protocol, supported the analysis of the data and the interpretation of the analyses and critically revised the paper. DC, MJ and MB helped to develop the protocol, interpret the findings and critically revised the paper. All authors approved the final version of the paper.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: David Currow and Miriam Johnson were Chief Investigators of four trials included in the analysis, however, were not involved in the data extraction or analysis.

ORCID iDs

Jamilla A Hussain https://orcid.org/0000-0002-3644-6480
Miriam J Johnson https://orcid.org/0000-0001-6204-9158
David C Currow https://orcid.org/0000-0003-1988-1250

Data management and sharing

Statistical data files, additional charts and graphs are available from the corresponding author on request. The corresponding author does not have the right to share the original trial data.

Supplemental material

Supplemental material for this article is available online.

References

1. Hussain JA, White IR, Langan D, et al. Missing data in randomized controlled trials testing palliative interventions pose a significant risk of bias and loss of power: a systematic review and meta-analyses. *J Clin Epidemiol* 2016; 74: 57–65.
2. National Research Council (US) Panel on Handling Missing Data in Clinical Trials. *The prevention and treatment of missing data in clinical trials*. Washington, DC: National Research Council, 2010.
3. European Medicines Agency. *European Medicines Agency guideline on missing data in confirmatory clinical trials*. London: EMA, 2010.
4. Brueton VC, Tierney J, Stenning S, et al. Strategies to improve retention in randomised trials. *Cochrane Database Syst Rev* 2013; 12: MR000032.
5. Riley RD, Lambert PC and Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010; 340: c221.
6. Kaasa S and Loge JH. Quality of life in palliative care: principles and practice. In: Cherny NI, Fallon M, Kaasa S, et al. (eds) *Oxford textbook of palliative medicine*. 4th ed. Oxford: Oxford University Press, 2015, pp. 443–461.
7. Centre for Reviews and Dissemination. About PROSPERO, http://www.crd.york.ac.uk/PROSPERO/ (2014, accessed December 2014).
8. Debray TP, Moons KG, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods* 2015; 6: 293–309.
9. Centre for Multilevel Modelling. Centre for multilevel modelling: training, http://www.bristol.ac.uk/cmm/learning/ (2002, accessed May 2016).
10. Hox JJ. Chapter 4.1 Analysis strategy. In: Hox JJ (ed.) *Multilevel analysis: techniques and applications*. 2nd ed. East Sussex: Routledge, 2010, pp. 49–54.
11. Steele F. Module 7: multilevel models for binary responses – concepts, https://www.cmm.bris.ac.uk/lemma/pluginfile. php/2281/mod_resource/content/1/mod-7-concepts.pdf (2009, accessed January 2016).
12. Goldstein H, Browne W and Rasbash J. Partitioning variation in multilevel models. *Unders Stat* 2002; 1: 223–231.
13. Rabe-Hesketh S and Skrondal A. *Multilevel and longitudinal modeling using Stata – vol 1: continuous responses*. 3rd ed. College Station, TX: Stata Press, 2012.
14. Hussain JA, White IR, Johnson MJ, et al. Development of guidelines to reduce, handle and report missing data in palliative care trials: a multi-stakeholder modified nominal group technique. *Paper under review, unpublished*, 2021.
15. Treweek S, Bevan S, Bower P, et al. Trial forge guidance 1: what is a study within a trial (SWAT)? *Trials* 2018; 19: 139.
16. Collins LM, Schafer JL and Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; 6: 330–351.
17. Hardt J, Herke M and Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med Res Methodol* 2012; 12: 184.
18. Oken MM, Creech RH, Tormey DC, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* 1982; 5: 649–655.
19. Rosenberg AR, Barton K, Junkins C, et al. Creating a resilient research program: lessons learned from a palliative care research laboratory. *J Pain Symptom Manag* 2020; 60: 857–865.
20. Preston NJ, Fayers P, Walters SJ, et al. Recommendations for managing missing data, attrition and response shift in palliative and end-of-life care research: part of the MORECare research method guidance on statistical issues. *Palliat Med* 2013; 27: 899–907.
21. Tierney JF, Vale C, Riley R, et al. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PLoS Med* 2015; 12: e1001855.
22. Medical Research Council. *Developing and evaluating complex interventions: new guidance*. London: MRC, 2008.
23. Riley RD, Lambert PC and Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010; 340: c221. DOI: 10.1136/bmj.c221.
24. Stewart LA, Clarke M, Rovers M, et al. & Group P-ID. Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD statement. *JAMA* 2015; 313: 1657–1665. DOI: 10.1001/jama.2015.3656.