Palliative and Supportive Care

cambridge.org/pax

A survey of statistical methods utilized for analysis of randomized controlled trials of behavioral interventions

Rebecca Tutino, M.A.[1,2], Elizabeth Schofield, DR.P.H.[1], Rebecca M. Saracino, PH.D.[1], Leah Walsh, M.A.[1,2], Emma Straus, M.P.H.[1] and Christian J. Nelson, PH.D.[1]

[1]Department of Psychiatry and Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, NY, USA and [2]Department of Psychology, Fordham University, Bronx, NY, USA

Original Article

Cite this article: Tutino R, Schofield E, Saracino RM, Walsh L, Straus E, Nelson CJ (2024). A survey of statistical methods utilized for analysis of randomized controlled trials of behavioral interventions. *Palliative and Supportive Care* **22**, 221–225. https://doi.org/10.1017/S1478951523000512

Received: 13 March 2023
Revised: 07 April 2023
Accepted: 11 April 2023

Keywords:
Statistics; Randomized control trial; ANOVA; hierarchical level modeling; Behavioral clinical trials

Corresponding author: Rebecca Tutino;
Email: tutinor@mskcc.org

Abstract

Objectives. Given the many statistical analysis options used for randomized controlled trials (RCTs) of behavioral interventions and the lack of clear guidance for analysis selection, the present study aimed to characterize the predominate statistical analyses utilized in RCTs in palliative care and behavioral research and to highlight the relative strengths and weaknesses of each of these methods as guidance for future researchers and reform.

Methods. All RCTs published between 2015 and 2021 were systematically extracted from 4 behavioral medicine journals and analyzed based on prespecified inclusion criteria. Two independent raters classified each of the manuscripts into 1 of 5 RCT analysis strategies.

Results. There was wide variation in the methods used. The 2 most prevalent analyses for RCTs were longitudinal modeling and analysis of covariance. Application of method varied significantly by sample size.

Significance of results. Each statistical analysis presents its own unique strengths and weaknesses. The information resulting from this research may prove helpful for researchers in palliative care and behavioral medicine in navigating the variety of statistical methods available. Future discussion around best practices in RCT analyses is warranted to compare the relative impact of interventions in a more standardized way.

Introduction

Randomized control trials (RCTs) are used widely across palliative care and behavioral medicine research to evaluate the efficacy of interventions. The defining features of an RCT are the presence of control groups and randomization, which remove allocation bias and minimize confounding effects. RCTs are thus considered the most rigorous way to determine that a cause–effect relationship exists between treatment and outcomes and are considered the “gold standard” as they allow for the minimization of biases introduced from confounding variables or covariates (Sibbald and Roland 1998).

Despite the prevalence of RCTs, it can be difficult to select a statistical method to analyze the results due to the range of available options without clear guidance. The assessment of change using a pretest–posttest control group design is a deceivingly straightforward task, as there is often a lack of clarity around when and how to use different methods (Rudestam and Newton 2012; Wilkinson 1999).

There have historically been 5 main ways to analyze continuous, individual-level RCT outcomes: analysis of variance (ANOVA) of change scores, ANOVA of follow-up scores, analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA), and longitudinal modeling, each with relative strengths and weaknesses. The first 2 options are the analysis of follow-up scores or the analysis of change scores, both of which are typically done using ANOVA or t-test of 2 groups. ANOVA is a method of separating variability on a dependent variable in order to test hypotheses regarding differences in means (Maxwell et al. 2017). Although these methods are straightforward, they assume that randomization balances the groups on pretest scores and do not adjust for baseline differences (see Table 1). It has been argued that these 2 methods of analysis should be avoided as they are often utilized inappropriately (Rudestam and Newton 2012). However, these methods are accepted among those who support the posttest-only method, which rationalizes that randomization should control for between-group baseline differences and that the inclusion of a pretest can reduce external validity by confounding change from an intervention (Vickers 2005b). Theoretically, if a study has a large enough sample, baseline characteristics including baseline measures should be balanced;

https://doi.org/10.1017/S1478951523000512 Published online by Cambridge University Press

however, determining how large a sample need be to achieve such balance may not be known, and for smaller samples, there may indeed be imbalance due to random chance that need be accounted for in the analysis.

ANCOVA (controlling for baseline score) is a method that adjusts for differences on the covariate by including the covariate (baseline score) as a continuous predictor variable in the analysis (Maxwell et al. 2017). This method thus accounts for baseline differences in primary outcome. ANCOVA also has higher statistical power compared to posttest and change score ANOVAs (Vickers 2001). Thus, this advantage may be particularly useful for analyzing studies with smaller sample sizes. ANCOVA can also be extended to incorporate time effects when using repeated measures and randomization strata as covariates (Vickers 2005a).

A MANOVA is a method of ANOVA that has 2 or more dependent variables (Warne 2014); the multiple dependent variables may be a single variable measured at multiple longitudinal timepoints. Similar to ANCOVA, this method accounts for baseline differences. However, it has been suggested that results using MANOVA are often misinterpreted, as the main effect is not the analysis of interest but rather the interaction effect (Vickers 2005b). There is also growing evidence to suggest that the field of psychology specifically is unfamiliar with the proper statistical procedures after rejecting a null hypothesis (Warne 2014).

Longitudinal modeling, such as mixed effects models with a random per-person intercept or generalized estimating equations that adjust variance estimates based on within-subject correlation, has also grown in popularity. Mixed effects models are regression models that explicitly incorporate a random per-person intercept to account for the within-person variation, while generalized estimating equations treat the per-person correlation as a nuisance variable and simply estimate the common correlation among data from the same person. These methods are often highly regarded as they can utilize all available data for participants lost to follow-up and can analyze multiple dependent variables. This method is also able to analyze unbalanced time points, a clear advantage to ANOVA. However, interpretation and choices of covariance structures and parameters may become overly technical.

Thus, there are many analysis options and a lack of clear guidance for selecting one method over another. This is especially true of RCTs determining the efficacy of behavioral interventions, also known as behavioral clinical trials, as the outcome variable of such studies is often continuous (Vickers 2005b). There has also been a growing appreciation of the differences in optimal methodology between behavioral clinical trials and standard pharmacological trials (Bacon et al. 2015; Penzien et al. 2005). Accordingly, behavioral clinical trials have added complexity in terms of research design and guidelines as well as noted limitations in their execution and dissemination (Bacon et al. 2015). While recommendations on the determination of sample sizes via statistical analyses have been outlined (Penzien et al. 2005), outcome analysis guidelines have yet to be established, despite the growing call for preregistration, which requires researchers to make a thoughtful selection of their statistical analysis plan a priori.

In response to the recent acknowledgment of the field's replicability crisis (Open Science Collaboration 2015), metascience has emerged as a scientific social movement that seeks to use quantification of science to diagnose issues in research practices with the goal of improving them (Peterson and Panofsky 2023). Metascience of statistical analyses may prove particularly useful, as Breznau et al. (2022) have noted idiosyncratic variation among researchers' analytic choices, even when working with the same data, and suggested that this may be especially true for behavioral research. Given the rise of palliative care and behavioral medicine interventions (e.g., Cognitive Behavioral Therapy, Motivational Interviewing, Meaning Centered Psychotherapy, etc.), it is critical to characterize the analytic patterns employed (Breitbart et al. 2018; Funderburk et al. 2018). Thus, the present study aimed to characterize and understand any patterns in the predominate methods utilized in recently published peer-reviewed RCTs in top palliative care and behavioral medicine journals with the goal of highlighting potential opportunities to reform future scientific practices.

Methods

Four journals with some of the most impactful research in palliative care and behavioral medicine and psycho-oncology were selected for analysis: *Annals of Behavioral Medicine, Health Psychology, Psycho-Oncology*, and *Psychosomatic Medicine*. These journals were selected based on study team consensus that they represent a sampling of (i.e., not intended to be exhaustive) some of the most widely respected journals in the field and the study team's interest in psycho-oncology. Inclusion criteria were (1) peer-reviewed publication in one of the 4 target journals; (2) RCT design with randomization at the participant level; and (3) analysis of the intervention effect on a continuous primary outcome. Studies where the primary outcome was feasibility (e.g., recruitment, retention, etc.) were excluded as these outcomes do not require inferential statistics. Studies were also excluded if they analyzed more than 2 follow-up time points as this design likely addresses questions beyond the scope of a pre–post analysis or if they were secondary analyses or an analysis only of mechanisms (i.e., moderation or mediation) and not reporting the main effect of the intervention.

IRB approval was not necessary for this review. An electronic query using the PubMed search engine was conducted for all manuscripts published in the 4 target journals during the calendar years 2015–2021. Articles were then excluded if they were not RCTs. Next, each article was deemed as eligible or ineligible based on the study inclusion criteria, resulting in the final analyzable manuscripts.

Among the manuscripts deemed eligible, 2 raters independently classified each study based on its statistical methods for the primary outcome prior to a consensus meeting of at least 3 raters where classifications were finalized after discussing any inter-rater disagreement. Classification categories were determined in concordance with Rudestam and Newton (2012), who delineated 4 primary methods for analyzing pre–post effects: (1) ANOVA of posttest scores, (2) ANOVA of change scores, (3) ANCOVA, and (4) MANOVA. In the general modeling context, these 4 methods translate to (1) regression of posttest scores without adjustment for pretest, (2) regression models of change scores without adjustment for pretest, (3) regression of posttest scores with adjustment for pretest, and (4) repeated measures ANOVA models with multiple observations per person, respectively. A fifth option, (5) multilevel modeling such as generalized estimating equations and hierarchical level modeling, was also included. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Descriptive statistics were calculated based on the classifications; frequencies were calculated for each method by journal, and overall classifications were compared by sample size using Kruskal–Wallis due to non-normality and by journal using Chi-square tests. For the Chi-square test of 5 methods among 4 journals,

https://doi.org/10.1017/S1478951523000512 Published online by Cambridge University Press

Table 1. Pros and cons of statistical approaches

Statistical approach	Outcome	Baseline adjustment?	Pros	Cons	Base SPSS code
ANOVA of follow-up scores	Y_2	No	- Easy to compute and for readers to understand	- Assumes that randomization balances the groups on pretest scores and does not adjust for baseline differences	ONEWAY *y2* BY *group*
ANOVA of change scores	$Y_2 - Y_1$	No	- Can identify high-change and low-change individuals	- Ceiling and floor effects may be masked	COMPUTE *d* = *y2* – *y1*
				- Assumes that randomization balances the groups on pretest scores and do not adjust for baseline differences	ONEWAY d BY *group*
ANCOVA	Y_2 or $Y_2 - Y_1$	Yes	- Generally has greater statistical power to detect a treatment effect (Wasserstein and Lazar 2016)		UNIANOVA *y2* BY *group* WITH *y1*
			- Accounts for baseline differences in primary outcome		
			- Can be extended to incorporate time effects (for repeated measures) and randomization strata as covariates (Vickers 2005a)		
MANOVA	Y_1 and Y_2	Yes	- Accounts for baseline differences in primary outcome	- Interaction term may be overlooked	GLM *y1 y2* BY *group*
				- Results are oftentimes misinterpreted	/WSFACTOR = time 2 Polynomial
					/WSDESIGN = time/ DESIGN = *group*.
Longitudinal modeling	Y_1 & Y_2	Yes	- Inclusion of baseline data and reduction of bias for participants who were lost to follow-up	- More complex, may be more difficult to interpret	MIXED *y* BY *group time*
			- Increases the effective sample size and statistical power		/FIXED = *group time group*time*| SSTYPE(3)
			- Allows researcher to identify time-point driving a specific effect (when multiple follow-ups)		/PRINT = G SOLUTION/METHOD = REML
					/RANDOM = INTERCEPT | SUBJECT(*id*) COVTYPE(UN).

Note. This table was adapted from Rudestam and Newton (Rudestam and Newton 2012). The syntax above assumes that the variables *y1* and *y2* are the baseline and first follow-up outcome measures, respectively, and *group* is the factor indicating group assignment. For the longitudinal model, the data need to be structured as one line per timepoint, indexed via the *time* and *id* values for timepoint and per-participant identifier, where *y* is the outcome at a given timepoint.

a sample of 183 manuscripts provides 80% power to detect a standardized effect size of at least Cohen's $w = 0.31$, a medium effect. Adjusted analysis was conducted via multinomial regression of the method on journal and log-transformed sample size, with an overall (type 3) test of the journal and sample size effects.

Results

The 7-year electronic query netted 3,989 manuscripts, of which 380 (10%) were identified as RCTs. Among all RCT manuscripts, 197 (52%) were excluded based on initial eligibility criteria as described above, resulting in 183 (48%) analyzable manuscripts from *Annals of Behavioral Medicine* (49 manuscripts), *Health Psychology* (41 manuscripts), *Psycho-Oncology* (58 manuscripts), and *Psychosomatic Medicine* (35 manuscripts). The consensus team classified all 183 manuscripts into one of the 5 distinct categories for statistical methods.

The most prevalent analytic method for the included RCTs was longitudinal modeling ($n = 58$, 32%), followed by ANCOVA controlling for baseline ($n = 42$, 23%) and MANOVA ($n = 40$, 22%). While longitudinal modeling (method 5) was the most prevalent method overall and for 3 of the individual journals, manuscripts in *Psychosomatic Medicine* more frequently used a MANOVA (method 4; 37%); however, this differential result was not statistically significant.

Sample size for the included studies ranged widely from 19 to 2,005 participants. Distributions of sample sizes, by method, are depicted in Figure 1. Statistical methods varied significantly by sample size ($p = 0.008$), such that manuscripts with larger sample sizes were more likely to employ ANOVA methods (of either change scores or follow-up scores, methods 1 and 2) and those with

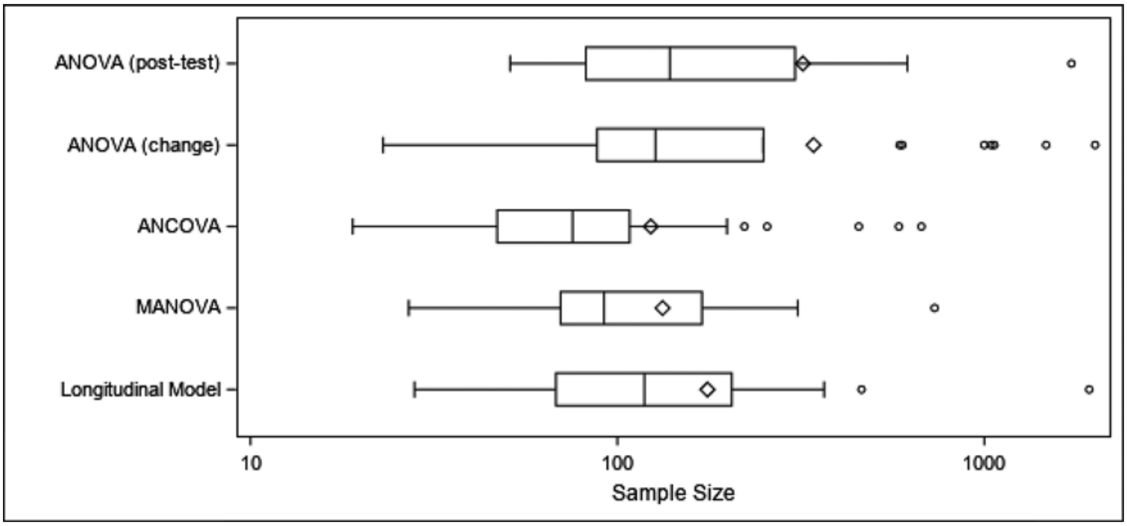

Figure 1. Distribution of Sample sizes per method.
Note. Sample size is depicted on a logarithmic scale.

smaller sample sizes were more like to use method 3 (ANCOVA). In a model including both log-transformed sample size and journal, only sample size was significantly associated with statistical method ($p = 0.03$).

Discussion

The great variability of analytic methods observed highlights the variety of options researchers have when selecting an analysis method, each with its own pros and cons (Table 1). Standardized guidelines outlining this decision-making process are of particular relevance given the growing utilization of registered reports, which require researchers to present their analysis plan a priori. As such, these guidelines would have the potential to aid in open science reform.

The prevailing method in these influential peer-reviewed palliative care and behavioral medicine journals was longitudinal modeling (method 5). As discussed in Table 1, an advantage of longitudinal (multilevel) modeling is the inclusion of baseline data for participants who were lost to follow-up. That is, multilevel model methods employ available-case analysis, whereas analyses of change scores or follow-up scores necessitate listwise deletion. Thus, multilevel models increase the effective sample size and statistical power for multilevel model methods compared to others and reduce bias related to participants lost to follow-up. If a researcher can assume attrition is random, listwise deletion is only a concern for power and not bias, but this assumption is rarely true in behavioral sciences.

Another advantage of longitudinal modeling is that one can extract multiple comparisons of interest (e.g., interval-specific time effects or pair-wise group comparisons) from the single model when appropriate contrasts are used. However, some recommend the use of multiple ANOVAs over longitudinal modeling as researchers may misuse this technique (Vickers 2005b). The biggest barrier to utilizing longitudinal modeling is a lack of familiarity with the techniques and computational logistics. For example, longitudinal data may need to be restructured to the less familiar "long" format where multiple observations per person are disaggregated into separate rows. This can be accomplished fairly succinctly using something like the CASESTOVARS function in SPSS but adds another layer of complexity if the researcher is not well versed in data management. Despite concerns that this technique may be misused, researchers may still select these more sophisticated analyses for perceived increase in publication potential.

The second most prevailing method in the 4 journals was ANCOVA controlling for baseline (method 3). As previously mentioned, one strength of this method is that it be extended to incorporate time effects (for repeated measures) and randomization strata as covariates, which has the benefit of potentially increasing power (Kalish and Begg 1985; Vickers 2005a). Another strength of this method is that it generally has greater statistical power to detect a treatment effect (Wasserstein and Lazar 2016), making it advantageous when analyzing smaller sample sizes, which may be more common in behavioral clinical trials. Accordingly, studies in the current review with smaller sample sizes were more likely to employ ANCOVAs.

The third most prevailing method in these journals was MANOVA. One possible strength of MANOVA over the previously described ANOVAs is that it adjusts for baseline differences. According to those who support the use of analyses such as MANOVA, even though RCTs ought to be balanced on baseline measures for adequately sized studies, a random imbalance or smaller studies may be better advised to allow for a baseline adjustment. Analyses using t-tests or ANOVA on either follow-up or change scores are also likely more accessible to readers with limited statistical training than methods such as MANOVA, and it has been suggested that results using MANOVA are often misinterpreted (Vickers 2005b).

Statistical methods also varied significantly by sample size across all journals. The observation that manuscripts using ANOVA of follow-up scores (method 1) had the largest sample sizes is appropriate, given the statistical principle that if there is a large enough sample size, baseline differences will be negligible due to randomization. The observation that studies with smaller sample sizes were more likely to employ ANCOVAs also aligns with the claim that ANCOVAs have greater statistical power than ANOVAs (Vickers 2001). Given that research suggests psychological research

https://doi.org/10.1017/S1478951523000512 Published online by Cambridge University Press

is oftentimes underpowered and sample sizes have not increased over time (Marszalek et al. 2011), the sample size should be considered when the researcher is selecting the appropriate method of analysis.

In addition to the descriptive findings of the current study, an incidental observation was that none of the papers reviewed utilized an initial Bayesian framework. The use of Bayesian models for estimation falls outside of traditional testing of a null hypothesis, instead resulting in estimation of parameters with a "highest density region" or a Bayes factor for model comparison. One study by Yeung et al. (2020) first utilized null hypothesis significance testing followed by post hoc Bayesian analyses to further analyze their data, perhaps representing an acknowledgment of the limitations of traditional null hypothesis significance testing. Recent statements on the obsolescence of traditional significance testing, made by such sources as the American Statistical Association (Wasserstein and Lazar 2016) and Nature (2019), have pointed to Bayesian methods and the utilization of the Bayes factor as an indicator of credibility of results. For analyses that utilize ANOVA or standard regression models, Bayesian methods have recently been made accessible and relatively user-friendly by incorporation into software such as SPSS (IBM Corp 2020).

In sum, the assessment of change using a pretest–posttest control method is a potentially complex task, and recent work has documented great variation among researchers' analytic choices (Breznau et al. 2021). While a statistical analysis plan should ultimately be driven by factors such as the research question, assumptions about the nature of change in the outcome, assumptions about attrition, and design factors including sample size, knowledge of the relative strengths and weaknesses of each common method used, as well as guidelines for their use may prove useful to researchers in palliative care and behavioral medicine. The information resulting from the current characterization of the literature and overview of the statistical methods available may help to inform this decision and aid in the development of future selection guidelines. Given the high levels of variability observed in this review, future discussion around best practices in RCT analyses is warranted to compare the relative impact of interventions in a more standardized way and aid in future scientific practice reform (e.g., preregistration, selection of an analysis plan a priori, open science reform, replicability, etc.).

Funding. Funding for this research was provided by the National Cancer Institute (T32CA009461-34; P30 CA08748-48; and K08CA25263).

Competing interests. The authors have no conflicts of interest to report.

References

Bacon SL, Lavoie KL, Ninot G, *et al.* Spring for the International Behavioural Trials Network (2015) An international perspective on improving the quality and potential of behavioral clinical trials. *Current Cardiovascular Reports* **9**(1), 1–6.

Breitbart W, Pessin H, Rosenfeld B, *et al.* (2018) Individual meaning-centered psychotherapy for the treatment of psychological and existential distress: A randomized controlled trial in patients with advanced cancer. *Cancer* **124**(15), 3231–3239.

Breznau N, Rinke EM, Wuttke A, *et al.* (2022) Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America* **119**(44) e2203150119.

Funderburk JS, Shepardson RL, Wray J, *et al.* (2018) Behavioral medicine interventions for adult primary care settings: A review. *Families, Systems, & Health* **36**(3), 368–399.

IBM Corp (2020) IBM SPSS statistics for windows (Version 27.0) [Computer software], Armonk, NY: IBM Corp.

Kalish LA and Begg CB (1985) Treatment allocation methods in clinical trials: A review. *Statistics in Medicine* **4**, 129–144.

Marszalek JM, Barber C, Kohlhart J, *et al.* (2011) Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills* **112**, 331–348.

Maxwell SE, Delaney HD and Kelley K (2017) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. New York: Psychology Press, Taylor & Francis Group.

Nature (2019) It's time to talk about ditching statistical significance. *Nature* **567**, 283.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* **349**(6251), aac4716 .

Penzien DB, Andrasik F, Freidenberg BM, *et al.* (2005) Guidelines for trials of behavioral treatments for recurrent headache: American Headache Society Behavioral Clinical Trials Workgroup. *Headache: The Journal of Head and Face Pain* **45**, S110–S32.

Peterson D and Panofsky A (2023) Metascience as a scientific social movement. *Minerva*, 1–28.

Rudestam K and Newton R (2012) *Statistical Consultant: Answers to Your Data Analysis Questions*. Thousand Oaks, CA: Sage Publications.

Sibbald B and Roland M (1998) Understanding controlled trials. Why are randomized controlled trials important? *BMJ* **316**, 201.

Vickers AJ (2001) The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Medical Research Methodology* **1**, 6.

Vickers AJ (2005a) Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology* **5**, 1–12.

Vickers AJ (2005b) Analysis of variance is easily misapplied in the analysis of randomized trials: A critique and discussion of alternative statistical approaches. *Psychosomatic Medicine* **67**, 652–655.

Warne RT (2014) A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research & Evaluation* **19** 1-10 .

Wasserstein RL and Lazar NA (2016) The ASA statement on p-values: Context, process, and purpose. *The American Statistician* **70**(2), 129–133.

Wilkinson L Task Force on Statistical Inference, American Psychological Association, Science Directorate (1999) Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* **54**(8), 594–604.

Yeung V, Sharpe L, Geers A, *et al.* (2020) Choice, expectations, and the placebo effect for sleep difficulty. *Annals of Behavioral Medicine* **54**, 94–107.

https://doi.org/10.1017/S1478951523000512 Published online by Cambridge University Press