ARTICLE IN PRESS

Social Science & Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Impact of missing data strategies in studies of parental employment and health: Missing items, missing waves, and missing mothers

Cattram D. Nguyen[a,b,c,*], Lyndall Strazdins[d], Jan M. Nicholson[a,b], Amanda R. Cooklin[a]

[a] Judith Lumley Centre, La Trobe University, Melbourne, Australia
[b] Murdoch Children's Research Institute, Melbourne Australia
[c] Department of Paediatrics, University of Melbourne, Melbourne, Australia
[d] National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia

ARTICLE INFO

Keywords:
Work-family conflict
Maternal mental health
Missing data
Non-response
Multiple imputation

ABSTRACT

Background: Understanding the long-term health effects of employment – a major social determinant – on population health is best understood via longitudinal cohort studies, yet missing data (attrition, item non-response) remain a ubiquitous challenge. Additionally, and unique to the work-family context, is the intermittent participation of parents, particularly mothers, in employment, yielding 'incomplete' data. Missing data are patterned by gender and social circumstances, and the extent and nature of resulting biases are unknown.
Method: This study investigates how estimates of the association between work-family conflict and mental health depend on the use of four different approaches to missing data treatment, each of which allows for progressive inclusion of more cases in the analyses. We used 5 waves of data from 4983 mothers participating in the Longitudinal Study of Australian Children.
Results: Only 23% had completely observed work-family conflict data across all waves. Participants with and without missing data differed such that complete cases were the most advantaged group. Comparison of the missing data treatments indicate the expected narrowing of confidence intervals when more sample were included. However, impact on the estimated strength of association varied by level of exposure: At the lower levels of work-family conflict, estimates strengthened (were larger); at higher levels they weakened (were smaller).
Conclusions: Our results suggest that inadequate handling of missing data in extant longitudinal studies of work-family conflict and mental health may have misestimated the adverse effects of work-family conflict, particularly for mothers. Considerable caution should be exercised in interpreting analyses that fail to explore and account for biases arising from missing data.

1. Introduction

Analyses of longitudinal cohort data are established as the 'gold standard' method to ascertain the long-term health effects of social determinants, such as employment, on participants' health and well-being over time (Thiese, 2014). Since the ground-breaking Whitehall studies (Marmot et al., 1978) longitudinal studies of work and health continued to build evidence about the ways in which unemployment, employment and employment conditions determine both physical and mental health and generate health inequalities (Berger et al., 2005; Cheng et al., 2000; de Lange et al., 2003; Dinh, Strazdins & Welsh, 2017; Dirlam and Zheng, 2017; Ferrie et al., 2002). The nature and quality of work is now established as a key social determinant of health for all adults globally (Marmot, 2005). Evidence is also emerging that another aspect of working life, *work-family conflict*, is also an important social determinant of health, particularly for parents (Amstad et al., 2011; Nohe et al., 2014). Work-family conflict (WFC) is defined as the conflict or strain (e.g., time, energy) that arises when demands of both work and home are incompatible (Greenhaus and Beutell, 1985). WFC has been associated with poorer mental health in parents, and poorer family functioning to a degree that measurably influences children's social and emotional well-being (Heinrich, 2014; Vieira et al., 2016). Like many employment-related determinants of health, the distribution of WFC conflict is socially patterned, including by gender and socio-economic status, as are the patterns of missing data (Cooklin et al., 2016; Rothenbühler and Voorpostel, 2016). The nature and extent of biases introduced into cohort studies via these patterns are poorly understood. The current study addresses this gap, using an illustrative example with missing data. We model the effects of a work-related health determinant (WFC), on a health outcome (mental health), for a

* Corresponding author. Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria 3052, Australia.
E-mail address: cattram.nguyen@mcri.edu.au (C.D. Nguyen).

https://doi.org/10.1016/j.socscimed.2018.03.009
Received 21 August 2017; Received in revised form 22 February 2018; Accepted 4 March 2018

group potentially vulnerable in the labour force – mothers of young children.

There are increasing calls in the WFC and health literature for more longitudinal research about the nature and trajectories of employment-linked determinants (Amstad et al., 2011; Nohe et al., 2014). Multiple waves of data are necessary to unravel the likely long-term adverse health effects and the sequence of causation. Longitudinal studies also provide valuable evidence about the mechanisms and pathways via which work-related exposures such as WFC can emerge, or be prevented. Through these studies, WFC can be tracked across different stages of the family life cycle, as can the inter-generational transmission of health impacts from parents to their children.

The issues of incomplete data, selective participant attrition, and the potential biases arising from these phenomena are key challenges in establishing the magnitude and social patterning of the health effects of WFC and many work-related variables. To an extent, these are issues for all longitudinal cohort studies – participants 'dropout' or withdraw from data collection, either entirely or from individual waves (i.e., wave non-response), or tender incomplete data in interviews or self-report questionnaires (i.e., item non-response). Against this background however, employment-dependent variables (e.g., WFC) and health outcomes (e.g., mental health) incur additional missing data complexities, which are unique to understanding work-health relationships. Under-employment, unemployment 'churning' and intermittent attachment to the workforce are markers of disadvantage and poorer earnings or career trajectories over time (Benach et al., 2014; Broom et al., 2006; Butterworth et al., 2013). These circumstances yield both 'missing' data about work exposures and poorer health outcomes.

For women, a key reason for under- or intermittent employment is caregiving and parenthood in particular. These key variables under investigation in studies of work and women's health similarly drive missing data. Workforce transitions are especially common for mothers across the childbearing years, in addition to other forms of gender-related disadvantage in the labour market. Mothers' absence (due to missing data) or under-representation in many analyses of work and health relationships obscures their particular work-health vulnerabilities (Campos-Serna et al., 2013; Mauno et al., 2012; Stier and Yaish, 2014). Mothers not employed for one wave (or more) will therefore 'skip' employment-related items in the data collection instrument (for any non-employed waves). 'Complete data', that is, data on all employment-related variables (including WFC) *for all waves* of data collection in a longitudinal study do not capture the reality of women's labour force participation. Analysing work and health data from those with only complete employment information effectively constrains the analytic sample to a subgroup who have participated *continuously* in work for the life of the study (or at least continuously at all data collection intervals). Inadvertently, this list-wise deletion practice can lead to longitudinal samples with unrepresentative characteristics, threatening external validity, and potentially biasing estimates of the relationship between the main variables of interest – work and mental health. Overall, complete-data analyses impede accurate analyses, in the present case for how work shapes health and health inequality, especially its gendered patterning.

What options are available for handling this complexity? What are the implications for study findings? How can we improve precision in estimating the longer-term or cumulative effects of workplace risk factors on parents' health, accounting for their intermittent workforce participation? We address these questions of missing cohort data by using an illustrative example focusing on one unique study population - mothers of young children - and on one work-related mental health determinant – work-family conflict.

1.1. Missing data mechanisms

In the literature on missing data, *missing data mechanisms* are commonly used as a framework for describing the processes underlying missing data (Little and Rubin, 2002; Sterne et al., 2009). They provide a means for articulating the assumptions made about the processes leading to the missing data, and for considering appropriate methods for handling the incomplete values. Values are said to be *missing completely at random* (MCAR) if the probability of missingness is not related to the values of the data themselves, observed or missing. Under MCAR, there are no systematic differences between the observed and missing data (Sterne et al., 2009). MCAR missingness could occur, for example, if there was a glitch with an online survey, which randomly led to missing WFC items for some participants. Alternatively, the mechanism is described as missing at random (MAR) if the probability of missingness is not related to the values of the missing data, *given (i.e., conditional on) the observed data*. For example, sole-parent families may be more likely to have missing WFC data. The mechanism is MAR if any systematic differences between the missing and observed data can be explained by the sole parenting status. So long as sole-parenting status is observed, then controlling for sole parenting can account for differences between those with and without missing data (Lang and Little, 2016; Sterne et al., 2009). The mechanism is missing not at random (MNAR) if the distribution of the missing data depends on the values of the missing data themselves. Say people with WFC are more likely to miss survey items (that ask about WFC) due to time constraints when juggling work and home commitment; then, participants with more missing data are also those with the highest WFC.

If there are systematic differences between participants with and without missing data, it suggests that the mechanism is not MCAR. Formal tests have also been developed for testing whether data are MCAR (Little, 1988). However, it is not possible to distinguish MAR from MNAR mechanisms without knowing the actual values of the missing data. The untestable nature of these assumptions is an inherent challenge with missing data analyses. In practice, assumptions about missing data mechanisms can be considered in relation to substantive knowledge about the possible reasons for the missing information.

1.2. Current approaches for handling missing data

Several statistical methods exist to handle missing data (Graham, 2012; Little and Rubin, 2002). The most common method is a *complete-case analysis*, which excludes participants with incomplete data on any variables in the analysis. Complete case analyses are convenient (and are the default method in many statistical packages); however, they are (generally) only valid under MCAR and can produce substantial bias if the complete cases are not representative of the entire sample (and if the analysis does not control for predictors of missingness). Discarding information from the incomplete cases can also lead to reduced power, compared to analyses that can utilise the partially observed data (Graham, 2009; Schafer and Graham, 2002). Other 'ad hoc' methods include single imputation methods that replace each missing value with a single imputed value (e.g., mean imputation, or last observation carried forward). In general, these methods are not recommended as they underestimate variance estimates and make strong assumptions (e.g., that a participant's outcome does not change following drop-out) (Little and Rubin, 2002).

'Principled' methods include multiple imputation, non-response weighting, and likelihood-based methods. These methods are increasingly being recommended by journals and reporting guidelines as preferred methods for handling missing data (Little et al., 2012). *Multiple imputation* (MI) replaces each missing value with multiple plausible values drawn from an imputation model to produce multiple completed datasets, and then proceeds with standard analysis methods applied to each to completed dataset. The multiple results are then combined using arithmetic rules to give an overall result with standard errors that account for the uncertainty of the imputed values (Little and Rubin, 2002). *Non-response weighting* attaches a weight to the complete cases to make them representative of the entire sample (Seaman and White, 2011). The weights are based on predicted probabilities of being a