International Journal of Nursing Studies 76 (2017) 36–44

Contents lists available at ScienceDirect

International Journal of Nursing Studies

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

Methodological approaches in analysing observational data: A practical example on how to address clustering and selection bias

Diana Trutschel[a,b,*], Rebecca Palm[a,c], Bernhard Holle[a], Michael Simon[d,e]

[a] German Center for Neurodegenerative Diseases (DZNE), Witten, Germany
[b] Martin-Luther-University Halle-Wittenberg, Halle/Saale, Germany
[c] University Witten/Herdecke, Witten, Germany
[d] University of Basel, Basel, Switzerland
[e] University Hospital Inselspital, Bern, Switzerland

ARTICLE INFO

Keywords:
Observational study
Nonexperimental studies
Health services research
Nursing research
Propensity score
Logistic models
Multiple logistic regression

ABSTRACT

Background: Because not every scientific question on effectiveness can be answered with randomised controlled trials, research methods that minimise bias in observational studies are required. Two major concerns influence the internal validity of effect estimates: selection bias and clustering. Hence, to reduce the bias of the effect estimates, more sophisticated statistical methods are needed.

Aim: To introduce statistical approaches such as propensity score matching and mixed models into representative real-world analysis and to conduct the implementation in statistical software R to reproduce the results. Additionally, the implementation in R is presented to allow the results to be reproduced.

Method: We perform a two-level analytic strategy to address the problems of bias and clustering: (i) generalised models with different abilities to adjust for dependencies are used to analyse binary data and (ii) the genetic matching and covariate adjustment methods are used to adjust for selection bias. Hence, we analyse the data from two population samples, the sample produced by the matching method and the full sample.

Results: The different analysis methods in this article present different results but still point in the same direction. In our example, the estimate of the probability of receiving a case conference is higher in the treatment group than in the control group. Both strategies, genetic matching and covariate adjustment, have their limitations but complement each other to provide the whole picture.

Conclusion: The statistical approaches were feasible for reducing bias but were nevertheless limited by the sample used. For each study and obtained sample, the pros and cons of the different methods have to be weighted.

What is already known about the topic?

Data in nursing health services research often is observational and clustered

Clustering and selection bias can lead to biased results

What this paper adds

The paper introduces common analytical strategies to address selection bias and clustering in observational research

Providing a vignette, researchers can replicate the used analytical strategies

1. Introduction

Nursing research aims to validate, refine and generate knowledge from studies that directly and indirectly affect the delivery of nursing care (Burns and Grove, 2009). Furthermore, evaluating health services, an aim of nursing research (AACN, 2015), requires research methods that achieve the highest internal validity possible to derive unbiased effect estimates of an intervention in a certain population in real-world settings. When threats to internal validity, such as selection bias or clustering, are not addressed through the study design, statistical methods are needed to reduce the bias of the effect estimates. Two major concerns influence the internal validity of effect estimates: selection bias and clustering. These two factors are the primary focus of this article.

We are motivated by our own observational study in health services research, in which three main data characteristics need to be addressed to find a suitable analysis method. Specifically, illustrated in Fig. 1, a *dichotomous* outcome of *clustered* data in a *observational* study was

* Corresponding author.
E-mail address: diana.trutschel@dzne.de (D. Trutschel).

http://dx.doi.org/10.1016/j.ijnurstu.2017.06.017
Received 6 December 2016; Received in revised form 12 June 2017; Accepted 29 June 2017

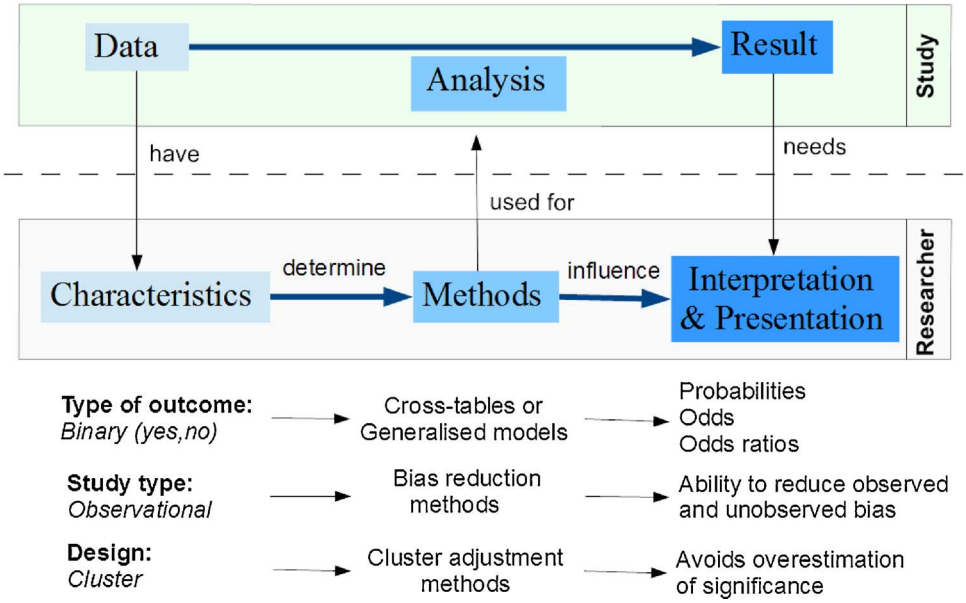

Fig. 1. In a study, the data analysis generates results. The data have their own characteristics, for example, a special outcome type, a unique study type or a specific design. These characteristics determine the choice between sophisticated methods for data analysis. Hence, the method directly influences the interpretation of the results and therefore must be carefully chosen using the skill of the researcher.

analysed.

First, the distribution of the outcome variable, which is one characteristic of our example data, influences the choice of the statistical method. Here, we analyse the use of case conferences as a binary outcome. Binary variables are summarised by probabilities, odds and odds ratios. A probability is defined as a relative frequency and can easily be understood (as a risk), whereas odds are an expression of relative probabilities – the ratio of the probability of the event occurring to the probability of no event occurring. Moreover, the odds ratio is the relation of two odds. However, because odds are not a probability, the interpretation is more difficult for practitioners (Greenland, 1987), and sometimes, odds are misinterpreted (O'Connor, 2013). Furthermore, if the model for effect estimation is not simple, then generalisable models that use link functions other than the identity functions are needed.

The second characteristic is the observational study type, which is used to collect data. In observational studies, the possibility of controlling factors that may influence the study outcome is limited to observed variables because randomisation is not part of the study design. Therefore, other options must be applied to reduce selection bias, which can contribute to over-/underestimations of the intervention effect (Starks et al., 2009). Hence, estimations of treatment effects through direct comparisons are prone to selection bias when the assignment to treatments is associated with the potential outcomes of the treatment (Ridder and Graeve, 2011).

Our example is an evaluation of special care units. Special care units serve dedicated patient populations that are in need of special care because of their health state. Special care units are implemented for conditions such as stroke, premature birth and dementia. For example, residents who reside in dementia special care units systematically differ from other residents because they are selected based on predefined criteria. Additionally, studies about dementia special care units typically have a multistage clustered data structure: residents are clustered within units, units are clustered within nursing homes, and nursing homes are clustered in provider systems. Selection bias may occur in every stage: residents in dementia special care units differ from residents in other care units, and nursing homes with dementia special care units may differ from nursing homes without dementia special care units.

Another problem that may arise in studies is the overestimation of how the significance of effects due to clustering influences the variance estimation of the effect. If more than one cluster is included in the study, a clustered or nested data structure is most likely present, and the error terms within a cluster are no longer independent. When the non-independence of the data is not accounted for in the statistical model, the odds for significant results increase. Hence, in our example, residents are clustered within nursing homes. This clustering must be considered when choosing the analysis method.

The nursing research literature contains many examples of observational studies that are necessitated to address selection bias and clustering. For instance, studies investigating the association of organisational characteristics, such as the work environment and patient or nurse outcomes, generally have to address both issues. For example, Zúñiga et al. (2015) explore the association between the work environment and care workers' perception of quality of care in 155 nursing homes in a cross-sectional study. To address selection bias, the authors employ a multilevel regression model with a range of variables as control factors (e.g., language region and unit size) and others as random effects (e.g., unit and hospital site) to address clustering.

In this article, we will introduce statistical approaches to reduce selection bias and clustering in a real-world data analysis example. We highlight the strengths and weaknesses of different methods, which are elucidated and discussed with respect to applying the methods to the chosen example study data. Additionally, we provide data and source code as a vignette (supplemental material) to show the practical implementation of the models separately and enable replicating the analysis with open-source software R (R Core Team, 2015), which might guide readers in applying the methods to their own studies and conditions.

Our aim here is not to provide a review of the methodological work within this field. Nevertheless, the following articles and books discussing propensity score (Austin et al., 2007; Belitser et al., 2011; Biondi-Zoccai et al., 2011; D'Agostino, 1998; Randolph et al., 2014; Sekhon, 2011; Stürmer et al., 2006), matching (Pimentel et al., 2015; Rosenbaum, 2002; Rubin, 2006; Stuart, 2010) and multivariate adjustment (Cepeda et al., 2003; Gelman and Hill, 2007) serve as guidance for our work.

The aim of this article is to highlight (1) why different methods should be used, (2) their application in a statistical software and (3) how to interpret the results produced by statistical methods.