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

* 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
0020-7489/ © 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

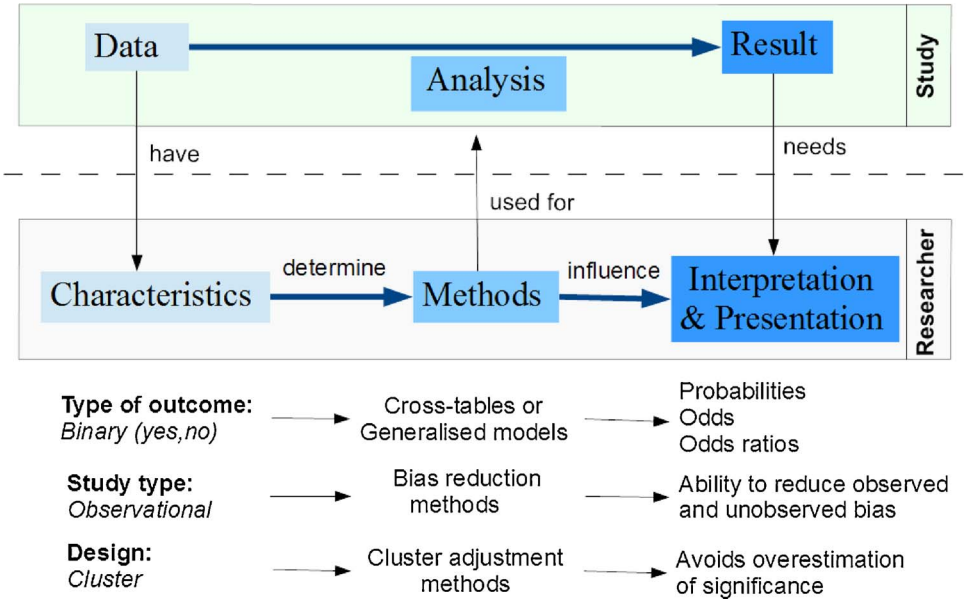

Fig. 1. In a study, the data analysis generates results. The data have their own characteristics, for example, a special outcome type, a unique study type or a specific design. These characteristics determine the choice between sophisticated methods for data analysis. Hence, the method directly influences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

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 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.

2. Materials and methods

2.1. Data and research example

The provided dataset is from the observational DemenzMonitor study (Palm et al., 2014, 2015). Data from 2013 were used for the analysis. The data consist of a convenience sample of 51 nursing homes, 109 care units and 1808 residents. After residents had been excluded due to only a two-group comparison being performed and predefined exclusion criteria, we used a dataset of n = 888 participants from 64 care units in 36 nursing homes (available in the supplemental material). Additionally, 53 residents with missing values in any of the variables were excluded. The primary question for this analysis was whether a dementia special care unit more frequently performs case conferences than traditional care units. The outcome variable was a binary indicator for whether the condition (1) was performed or not (0). Because the study used an observational design, residents in special care units and traditional units did not necessarily share the same characteristics, thus requiring an analytical approach to address selection bias. Furthermore, the clustering of residents in nursing homes leads to non-independent observations, again requiring an analytical approach that takes this clustering into account.

2.2. Procedure

Table 1 shows the two levels of analytical strategies for addressing the problem of unequal distributions of characteristics in the condition and comparison groups and the problem of clustering: (i) different models with different abilities to adjust for dependencies to analyse binary data and (ii) different methods to adjust for selection bias. Here, both analytical problems are addressed and combined in the analysis.

We distinguish two models for obtaining inference from the binary data: a crude model and a generalised linear mixed model. In the crude model, the results are not adjusted for the hierarchical data structure (clustering) or for differences in baseline characteristics, resulting in a higher risk of false-positive results. The generalised linear mixed model, which is a multilevel model without any additional control variables, addresses the clustering issue but does not address selection bias. We describe two methods for bias reduction that can be used for analysing data with dichotomous outcomes (by crude or advance model): (1) genetic matching on samples and (2) adjustment via the common regression model. All steps of this procedure, which are shown in Table 1, can be followed and adapted for other data sets using the provided Vignette (supplemental material), which shows the implementation with the programming language R (R Core Team, 2015). In this article, we will first introduce the crude model, then adjust for clustered data with the generalised linear mixed model, and finally use this model with all methods for bias reduction (only the shaded areas in Table 1).

2.3. Different models and their ability to adjust for dependencies

In our example, because we analyse a binary outcome variable, common methods for normally distributed variables and statistical tests such as Student's t-test and ANOVA cannot be used. Testing the differences between groups is similar to testing the differences between proportions in a contingency table, which refers to the 'crude model'. Testing the association between a dependent variable and a group of independent variables for a binary outcome requires a logistic regression model. The crude model is identical to a regression model with the group assignment (condition, control) as an independent variable without covariate adjustment. When observations are not independent, e.g., because of clustering in different nursing homes, a generalised linear mixed model is used.

2.3.1. Crude model

The crude model is a simple contingency table (upper part of Table 2) that provides an initial overview of the two-dimensional frequency distribution of cross-tabulated data – the distribution of a binary outcome variable (here, the performance of case conferences). From this table, probabilities and odds can be calculated (bottom part of Table 2). Hence, the (estimated) probability of an 'event will take place' can be calculated as a proportion from the frequencies in each group (see the supplemental material for equations), and differences in (estimated) probabilities between the two groups can easily be calculated. The probability of an event in a specific group is also known as risk; therefore, the risk ratio compares the probability of an event in one group to that in another (here, for example, the treatment group versus the control group). Often, the chance that something will occur is described as the odds (see the supplemental material for the equation). Although the interpretation is more difficult for practitioners (Greenland, 1987) because an odd is not a probability and sometimes is misinterpreted as a risk (O'Connor, 2013), the provided scale is indefinite and hence provides possibilities of working with other mathematical methods. The odds are the ratio of both probabilities, namely, the probability of an 'event will take place' versus the probability of an 'event will not take place' $p/(1-p)$, and it compares how much larger one probability is relative to another in a specific group. In our case, the (estimated) probability that a case conference was conducted in the control group was 0.8, and the (estimated) probability that a case conference was not conducted was 0.2. Hence, within the control group, the (estimated) probability that a case conference was conducted is four times higher than not, which indicates an odds of 4 (0.8/0.2). The widely used odds ratio is thus the ratio of both odds, namely, the odds of the treatment group related to the odds of the control group (see the supplemental material for equations). The odds ratio compares the difference in the odds between the two groups. If the odds are equal in both groups, then the odds ratio is equal to one. In our case, the odds ratio is 2.58, which means that the odds of receiving a case conference in the condition are higher than those in the control group.

Table 1
Different analytical strategies for selection bias reduction and/or cluster adjustment. The crude model, generalised linear model (GLM), is not able to adjust for clustered data or reduce bias in observational studies. A generalised linear mixed model (GLMM) is essential to account for multilevel data. Selection bias can be reduced by (1) including covariates in the regression model or (2) using a matching algorithm to reach a balance on the covariates between the investigated groups.

		Bias reduction	
		no	yes
Cluster adjustment	no	Crude (GLM)	GLM + Genetic matching GLM + Covariate adjustment
	yes	GLMM	GLMM + Genetic matching GLMM + Covariate adjustment

Table 2
Upper: a contingency table of a two-group comparison for a dichotomous outcome variable, where n_{ij} is the absolute amount of outcome i in group j. Lower: parameters, their estimates calculated from the contingency table and their interpretation.

		Group		
		Treat	Control	Marginal
Outcome	No	n_{11} (22)	n_{12} (119)	$n_{1.} = n_{11} + n_{1.}$ (141)
	Yes	n_{21} (224)	n_{22} (470)	$n_{2.} = n_{21} + n_{2.}$ (694)
	Marginal	$n_{.1} = n_{11} + n_{21}$ (246)	$n_{.2} = n_{12} + n_{22}$ (589)	$N = n_{.1} + n_{.2} = n_{1.} + n_{2.}$ (835)
Interpretation	Probabilities	$p_{\text{Treat}} = \frac{n_{21}}{n_{.1}}$ (0.91)	$p_{\text{Control}} = \frac{n_{22}}{n_{.2}}$ (0.8)	(Risk)Diff. $= p_{\text{Treat}} - p_{\text{Control}}$ (0.11)
	Odds	$\text{Odd}_{\text{Treat}} = \frac{p_{\text{Treat}}}{1-p_{\text{Treat}}}$ (10.18)	$\text{Odd}_{\text{Control}} = \frac{p_{\text{Control}}}{1-p_{\text{Control}}}$ (3.95)	Odds Ratio $= \frac{\text{Odd}_{\text{Treat}}}{\text{Odd}_{\text{Control}}}$ (2.58)

Table 3
Parameters of generalised linear model with the full sample and their interpretation. Each parameter $x \in (\beta_0, \beta_1, \beta_0 + \beta_1)$ of a generalized linear model with the form of $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ can be extracted and interpreted. The left column shows the estimates of the model and are interpreted as logarithmic odds, the middle column provides a transformation into (estimated) odds, and the right provides a transformation into probabilities.

Parameter	Transformation		
	No	$e^{\text{parameter}}$	$\frac{e^{\text{parameter}}}{1+e^{\text{parameter}}}$
β_0	$\text{logOdd}_{\text{Control}} = 1.37$	$\text{Odd}_{\text{Control}} = e^{1.37} = 3.95$	$p_{\text{Control}} = \frac{e^{1.37}}{1+e^{1.37}} = 0.8$
$\beta_0 + \beta_1$	$\text{logOdd}_{\text{Treat}} = 2.32$	$\text{Odd}_{\text{Treat}} = e^{2.32} = 10.18$	$p_{\text{Treat}} = \frac{e^{2.32}}{1+e^{2.32}} = 0.91$
β_1	logOdds ratio $= 0.95$	Odds ratio $= e^{0.95} = 2.58$	

The crude model provides the same results as the logistic regression model with only one independent variable for group assignment (case versus control). The logistic regression belongs to the family of generalised linear models (GLMs), which can handle different distributions of outcome variables. Assuming that $p = P(\text{Event} = \text{yes}|X)$ is the probability that an event occurs given the predictor variables X and that p_i is this probability for one response i, the generalised linear model adapts the linear relationship using the logit function (see the supplemental material for equations). The logistic regression model with a single dichotomous predictor variable for the assignment to the group 'Treatment' for a binary outcome (e.g., whether a case conference was performed) is a simple example. With the maximum likelihood method, the parameters of the logistic regression model, β_0 and β_1 here, can be estimated, and then, the inverse function of the logit $logit^{-1}(x) = e^x/(1 + e^x)$ provides the ability to assess the probability values of [0,1] (see the supplemental material for equations). In our case, this means that the estimated log odds and log odds ratios can support the (estimated) probabilities of receiving a case conference for each group.

In Table 3, the expressions of the estimated parameters, i.e., the estimates of the logarithmic odds and odds ratio, are listed (equations are explained in supplemental material). The exponentiated expressions of the model parameters are the odds and the odds ratio. The (estimated) probabilities of success in both groups are also given. For example, the parameter β_0 of the logistic regression model expresses the logarithmic odds of the control group, and the exponentiated value e^{β_0} is the corresponding odds, which indicates the chance that an 'event will take place' versus the opposite chance in the control group. The probability that an 'event will take place' in the control group is calculated using $e^{\beta_0}/(1 + e^{\beta_0})$. Due to the circumstances, the linear function can also be from a different family of functions; this type of model specification is the generalised linear model. Furthermore, some statistical programs provide the converted estimated values from such generalised models coincidently, but in R, these values must be calculated manually. Hence, the mathematical link and inverse link function (as shown in Table 3) must be known to provide the estimates in the required scale ((estimated) odds or probabilities in this case).

As with all regression models, the logistic regression can adjust for measured group differences (e.g., age or severity) when a binary outcome is predicted from a set of variables. Hence, proportions based on a dichotomous event are analysed using this widely used method (Ostir and Uchida, 2000).

2.3.2. Generalised linear mixed model

In our case, observations were collected from participants in care units nested in nursing homes. Therefore, one of the key assumptions of the logistic regression model – independence of observations – is violated. Because more than a half of the nursing homes (20 of 36) provided only one care unit, we use nursing home as only a cluster level (for more detail, a histogram of the number of participants in each care unit within the nursing homes is given in Fig. 2, Supplemental I). Hence, in this situation, the treatment is assigned at the individual level. In our example, the intra-class correlation coefficient (ICC) is 0.48 [0.2, 0.73], which means that 48% of the variation is explained by the variation between the nursing homes. Although a range of different estimators exists (see Wu et al., 2012 for details) [2012], here, we use the Fleiss-Cuzick estimator given by Zou and Donner (2004) to calculate the intra-class correlation coefficient on the proportional scale (see the Vignette for calculations; additionally, a model-based estimator is given).

Violation of this assumption of independence due to clustered data can lead to committing type I errors, e.g., finding an association where there is none. A solution to this problem is to apply a generalised linear mixed model. Generalised linear mixed models are an extension of the generalised linear models and are well established (see (Hardin and Hilbe, 2012; Stroup, 2012). They combine two statistical concepts: using linear mixed models to include random effects and using generalised models to model non-normal data. Hence, error terms that correspond to the different sources of variation in the data are added to the logistic regression model (Gelman and Hill, 2007), and the residual variance can be separated into components of the different involved levels (Li et al., 2011). In our example, the individual probability being statistically dependent on the nursing home where a participant lives is considered, and the variation between nursing homes and participants is quantified.

2.4. Methods for selection bias adjustment

In this section, we will introduce two basic approaches to address

selection bias: (1) matching and (2) regression, and provide a very brief description for why we take this approach. Because both approaches can use all covariates or the propensity score for adjustment, we provide the definition of the propensity score first.

2.4.1. Propensity score

The propensity score was introduced by Rosenbaum and Rubin (1983) and is defined as the conditional probability of being treated given a set of covariates. The definition of the propensity score for a single subject i (Eq. (1)) is the conditional probability of assignment to the treatment group ($Z_i = 1$) given a vector of observed covariates (x_i), where Z_i is assumed to be independent. Based on the observed control variables for each subject, a propensity score for membership in the treatment group is calculated from a logistic regression. Hence, the propensity score summarises different confounding factors into one dimension and can thus be used to achieve balance (Biondi-Zoccai et al., 2011) through adjustment methods, such as matching or regression models (D'Agostino, 1998). Balancing in this context means that the baseline characteristics in the treatment and control groups are the same (matching) or that balance differences are taken into account (regression). Using propensity score methods allows estimation of unbiased treatment effects if there is no unmeasured confounder (Williamson et al., 2012). Numerous literature reports that consider the impact of the selection of the model for propensity score estimation on the ability to reduce bias through the outcome model and also balance checks after application of the propensity score are available (Arpino and Mealli, 2011; Austin et al., 2007; Belitser et al., 2011; Leyrat et al., 2014; Nguyen et al., 2017; Rosenbaum and Rubin, 1983; Stürmer et al., 2006; Williamson et al., 2012).

$$\text{Propensity score} = e(x_i) = P(Z_i = 1 | X_i = x_i) \tag{1}$$

2.4.2. Genetic matching and hidden bias assessment

The matching of similar individuals in the treatment group with individuals from the control group, at least theoretically, is a frequently suggested approach for balancing observed control variables in both groups (Baser, 2006). The propensity score, confounding covariates or both can be used to match members of the condition and the control group to achieve covariate balance in both groups (Sekhon, 2011). Although this approach is theoretically appealing, in practice, balance is difficult to obtain, and researchers must repeatedly specify the propensity score model to approximate covariate balance between groups (Austin, 2009). Subsequently, several balance measurements for checking before and after matching have become available (Belitser et al., 2011).

Guidance on the use of matching is given by Stuart (2010), where the different available parameters to reduce bias due to covariates by choosing well-matched samples is explained. For example, matching can be done with replacement, which means that the controls can be used as a match multiple times. If the inclusion of multiple matched control observations because one treated observation matches more than one control observation is allowed, then ties have to be handled. Furthermore, whether an exact match is required or a defined distance between individuals is possible can be specified. For the matching problem, Sekhon (2011) proposed a genetic matching algorithm that automatically maximises covariance balance.

After matching, the average treatment effect $\hat{\Theta}$ can then be estimated from the matched sample in an unbiased manner under the assumption of there being no unobserved confounder by the difference in the means of the outcomes between both groups. Eq. (2) shows that for our example, the estimated average treatment effect (provided by the R-package for matching the type of estimand that can be specified) is equal to the difference in the (estimated) probabilities (or proportions) of the 'event' between the treatment group p_{Treat} and the control group p_{Control} from the contingency table (Section 2.3.1, Table 2).

Whereas matching can address only the balance of observed variables, researchers are also interested in what the effect of unobserved variables ('unobservables') might have been. Unobservables are the key advantage of randomisation in trials because with increasing sample size, randomisation automatically balances observed and unobserved covariates. The Rosenbaum bounds are used to test the robustness of conclusions to hidden biases from unobserved confounders (Rosenbaum and Rubin, 1983). The value Γ is the odds ratio of its effect on treatment assignment – i.e., how much an unmeasured confounder would increase the odds of the measured outcome.

The use of the Rosenbaum bounds requires independent and identical distributed data. In our example, there is a lack of this independence assumption. On the one hand, within multilevel data, observations within a cluster are not independent. On the other hand, matching with replacement may result in multiple uses of controls for different match units. Hence, a more modern method is required to handle such data assessing hidden bias (Zubizarreta and Keele, 2014). However, the genetic matching approach not only provides the advantage of reducing selection bias and being able to model the propensity score without specifying an outcome but also provides a means to assess 'hidden bias' from unobserved confounders.

After applying the genetic matching approach, the matched sample can then be further analysed, e.g., using generalised linear mixed models to adjust for clustering. Because of matching, the sample size is reduced and may reduce the power for the interested effect size estimation. However, with regression analysis, multiple effects are estimated with increasing requirements per degree of freedom. Matching avoids this problem because only the effect of interest has to be estimated. Nevertheless, matching may produce data with additional 'non-independent' observations, which then should be considered through analysis.

2.4.3. Covariate adjustment

The most common method for reducing selection bias is likely the inclusion of independent variables (covariates) in a multiple logistic regression model for dichotomous outcomes. Hence, analysis and bias adjustment are not separated. Including covariates within the regression model subsequently provides a conditional estimate of the treatment effect (given levels of the covariates), which could differ from the marginal effects. Therefore, the estimated coefficient from the model should be interpreted with caution.

The researchers are responsible for which covariates are considered to include into the model. One possibility is to use all suspected covariates that are relevant, but covariate adjustment methods are often limited in the possible number of covariates (D'Agostino, 1998), and if models include too many variables, they may fail to converge. Convergence failure in this context means that the model cannot be estimated computationally. An alternative approach to account for different covariates is to include the propensity score in the regression equation. This idea follows the same principles as outlined earlier but without conducting a matching procedure based on the propensity score. Instead, the propensity score is included as an additional covariate in the regression model. The new variable can then be included in the regression model as one covariate rather than as an amount of covariates to control for bias and to increase the precision of the treatment effect estimate. Including one or many variables decreases the sample size for each 'cell'; thus, models including more variables have a higher risk of non-convergence. Nevertheless, this method only adjusts for bias through a regression model (not independent from the outcome), and no hidden bias assessment is possible.

$$\begin{aligned}\hat{\Theta} &= \hat{\mu}_{\text{Treat}} - \hat{\mu}_{\text{Control}} = \frac{1}{N_{\text{Treat}}} \sum_i Y_{\text{Treat},i} - \frac{1}{N_{\text{Control}}} \sum_i Y_{\text{Control},i} \\ &= \frac{\text{no(Event=yes | Group=Treat)}}{\text{no(Group=Treat)}} - \frac{\text{no(Event=yes | Group=Control)}}{\text{no(Group=Control)}} \\ &= \text{P(Event=yes|Group=Treat)} \\ &\quad - \text{P(Event=yes|Group=Control)} \\ &\Leftrightarrow \hat{\Theta} = p_{\text{Treat}} - p_{\text{Control}}\end{aligned} \tag{2}$$

In summary. The two methods to adjust for selection bias introduced here are matching using the genetic algorithm and adjustment within the

regression model estimation. The first balances the sample independent from the outcome and provides a means of assessing 'hidden bias' from unobserved confounders. The other method adjusts for selection bias by estimating the contribution of each variable to the outcome within a regression framework. However, adding more variables can decrease statistical power in small samples because it increases the variance around the regression estimate by decreasing the number of degrees of freedom (Starks et al., 2009). Hence, for both matching and regression, the propensity score alone or in tandem can be used to achieve balanced samples. A combination of propensity score adjustment for a subset of covariates and covariate adjustment for the other is also possible.

2.5. Estimation of treatment effect

Although the parameter of interest is the average treatment effect during the analysis of our example study with binary outcomes, it corresponds here to the odds ratio or risk difference. Austin (2007) discussed different estimation methods in addition to the crude model and other propensity score methods being needed to assess the average treatment effect. These suggested methods have substituted using several covariates for using only the propensity score.

2.5.1. After matching

In addition to the marginal odds being calculated directly from the contingency table of the matched sample (for example, after propensity score matching), another possible method is model based. A logistic regression model with only one predictor variable for the assignment to the treatment group is fitted on the matched sample to estimate the impact of the treatment on the change in the odds of the outcome. This is also possible for the mixed model variant. The exponential parameter e^{β_1} from this model (Table 3) is therefore an estimate of the marginal odds ratio.

2.5.2. After covariate adjustment

The logistic regression model, which includes several covariates to adjust for their imbalance, provides only a conditional estimate of the treatment effect (by transformation of the coefficients as previously described), and the interpretation is in terms of adjusted changes in the corresponding covariates. Hence, the average treatment effect is available as the odds ratio/risk difference marginalised over the distribution of the included covariates. Therefore, the predicted probabilities for each individual given the confounders (sample data) are estimated under the treatment condition and under the control condition. The calculated mean probabilities $\overline{p}_{\text{treat}}$ and $\overline{p}_{\text{control}}$ can then be used to provide an estimate of the marginal odds ratio using $\frac{\overline{p}_{\text{treat}}}{1-\overline{p}_{\text{treat}}}\frac{1-\overline{p}_{\text{control}}}{\overline{p}_{\text{control}}}$ and the marginal risk difference by $\overline{p}_{\text{treat}} - \overline{p}_{\text{control}}$ (see also in the Vignette). Therefore, we use the logistic regression mixed model for our multilevel data.

3. Results

3.1. Crude model

According to the crude model (Table 2), 91% ($n = 224$) of residents in dementia special care units received a case conference, whereas only 80% ($n = 470$) in traditional care units received a case conference. The substantive interpretation would be that a patient is more likely to receive a case conference in dementia special care units than in traditional care units. Using the base logistic regression model with one binary predictor (dementia specific care unit or traditional care unit), i.e., the estimated model parameters β_0 (Intercept) or β_1 (dementia specific care units: treat), indicates that the model specification is not substantively different from the crude model. Hence, the retransformed values are equal to the estimates of the crude model. Table 3 shows how to obtain the transformed estimates of the logistic regression model estimates of Table 3, which were calculated as explained in Section 2.3.1.

The table shows that with the logistic regression model, the odds of obtaining the condition, i.e., the (estimated) probability of receiving a case conference versus not, is 3.95 in the control group and 10.18 in the treatment group. Hence, this results in an odds ratio of 2.58, which indicates that the odds of receiving a case conference is more than two and a half times higher in the treatment group than in the control group; in other words, being in the treatment group (relative to the control group) raises the odds of receiving a case conference. Using the inverse logit function, we can also provide the estimates in terms of probabilities (right column of Table 3) of receiving a case conference: 91% receive case conferences in the treatment group compared to 80% in the control group. These values indicate that an additional 11.3% receive a case conference in dementia special care units than in traditional care units. Table 1 in supplemental material also shows that the confidence interval of the odds ratio (corresponding to β_1) does not include 1. Hence, the difference of 11.3% between the two groups is assumed not to be random. We conclude that dementia special care units more often provide case conferences than traditional care units.

3.2. Generalised linear mixed model and adjustment methods for bias reduction

The results of the generalised linear mixed model are presented in Table 2 in the supplemental material. The odds ratio of 8.23 (see also in Table 4) is more than three times higher than the odds ratio in the generalised linear model ignoring the clustered data. However, the confidence intervals also increase (see the estimated confidence intervals in the Vignette), thereby increasing the p values due to the odds ratio with a covariance structure reflecting the dependencies of the observations. Although the precision of the estimates decreases, which may result from convergence problems with the estimation approach, adjustment is necessary to ensure that we do not overestimate our results.

3.2.1. Propensity score estimation

To address selection bias, we estimate the propensity scores for each observation. We model the group assignment using an additional generalised linear model that includes all individual related covariates (for general and health-related characteristics, see (Palm et al., 2014)) as fixed effects (no interaction terms). Fig. 3 in the supplemental material shows the unequal distributions of the estimated propensity scores between both groups and reflects the need for covariate adjustment to address selection bias.

Additionally, to account for the nested structure of the data and thus adjust for potential cluster-level unobserved confounders, we estimate the propensity scores using a generalised linear regression mixed model that includes the nursing homes as random effects (see the Vignette). However, this model failed to converge, and these estimates of the propensity scores could not be applied for further analysis, although Arpino and Mealli (2011) recommended their use as the matching variable in such multilevel settings, where the treatment is assigned at the individual level.

3.2.2. Genetic matching and hidden bias assessment

In this section, we show that either (1) all covariates or (2) only the propensity score as a summary of the covariates can be used for the genetic matching approach to balance the two groups. Furthermore, the matching quality due to hidden bias from unmeasured variables after matching is examined using the Rosenbaum bounds.

Initially, we used the genetic matching approach to determine the optimal covariate balance in the matched sample permitting replacements. The choice of the specific variables was based on theoretical considerations and generating a balanced sample in terms of general and health-related characteristics (see Palm et al., 2014). The algorithm samples a subset of 246 observations from each group out of the original sample, which is limited by the number of observations within the treatment group. For the outcome of interest, the estimated average treatment effect, the estimated average causal effect, and hence the (estimated) difference in probabilities between the two groups are 0.13.

This value corresponds to an estimated treatment effect using the crude model on the subset of the data received from matching. According to a simple *t*-test, this difference in means is assumed to be significant. To implement a balance check after matching, the Vignette shows a variety of univariate standardised statistics being employed for each covariate proposed by Sekhon (2011), and the result shows (Vignette) that for all given covariates, balance is achieved by matching.

Using the Rosenbaum bounds at this point provides the opportunity to assess the matching quality due to hidden bias from unmeasured variables. Unfortunately, the significant *p*-value upper bound of 0.05 will be exceeded by a hidden variable with a Γ of 1.2. If we allow a *p*-value upper bound of 0.1, then it will be exceeded by a hidden variable with a Γ of only 1.3, indicating that an unobserved covariate that produces only a 1.2–1.3-fold increase in the odds of the group assignment would change the *p*-value to non-significance. Therefore, we would conclude that the matched sample is sensitive to hidden bias.

Nevertheless, we intend to use this matched sample for further analysis to compare it with the results of the unmatched sample, which is definitely biased. For the matched sample, we used the same basic generalised linear mixed model with one binary factor (dementia specific care unit or traditional care unit) that we used for the unmatched sample.

Table 4 shows the results interpreted in terms of (estimated) odds (left columns) and probabilities (middle columns) for the model adjusted for the clustered structure and adjusted for selection bias using the matched sample (third row). The odds ratio of 3.9 is half the odds ratio estimated with the same model from the unmatched sample (Table 4). The null hypothesis of no difference in the use of the conditions between the two groups would not be rejected.

Second, we use the genetic matching approach with the propensity score as the only covariate to determine the optimal propensity score balance. To check the balance after matching, the overlapping coefficient of the propensity score proposed by Belitser et al. (2011) is implemented in the provided Vignette, and the results show (supplemental material, Fig. 4) that matching based on propensity score was successful. Using the Rosenbaum bounds to check the matching quality for unobserved variables results in the significant *p*-value upper bound of 0.05 (0.1) being exceeded by a hidden variable with a Γ of only 1.8 (2). The estimates of the odds differ (Table 4), e.g., the odds of receiving a case conference in the treatment group, which for the propensity score-adjusted model are more than twice the odds from the covariate-adjusted model (46.78 vs. 40.33), and the estimated odds ratios (4.81 vs. 3.9) for both models. However, the null hypothesis of no difference between the two groups cannot be rejected.

Since Pimentel et al. (2015) and Zubizarreta and Keele (2014), accounting for multilevel structure within the matching process and balancing checks after matching are possible. These modern methods should be considered in the future when analysing multilevel data from observational studies.

3.2.3. Covariate adjustment

Including other relevant covariates as fixed effects instead of using the propensity score as the single indicator of group assignment within the effect estimation model to adjust for selection bias is possible. However, non-convergence occurs when there is too little data for the number of parameters or when the proposed model is not suitable for the given data. Hence, the choice of which variables should be included in the model can be based on the degree of significance in the difference between the treatment and control groups in the baseline analysis. Furthermore, variables could be included as fixed or random effects as long as the model converges. Here, we include three variables as fixed effects and two as random effects.

Nevertheless, the model needs a considerable amount of computation time to estimate the parameters and confidence intervals due to the number of included variables. The results presented in Table 4 (bottom row) show that the odds ratio of 6.99 between the two groups is not significant.

Rather than all covariates being used as fixed effects, it is possible to include only the propensity score, a continuous variable, as the single indicator of group assignment in the generalised linear mixed model in order to adjust for selection bias. The results show that the odds ratio of 6.3 between the two groups given a fixed PS value is not significant (Table 4).

The additionally estimated marginal risk differences of the generalized linear mixed model including the propensity score (in brackets, Table 4) are of comparable size to that provided by the basic generalised linear mixed model.

4. Discussion and conclusion

4.1. In summary

In this article, we, with the aid of a real study example, illustrate different methods to analyse data with selection bias and clustering and with a dichotomous outcome. Additionally, we provide a vignette as the supplementary material to enable readers to follow a full analysis of this study example in *R* and to adapt this method for other studies. For our study example, Table 4 presents the results for all models and methods and highlights the marked difference between the applied methods and the computed estimates. For our example, addressing the dependencies with a mixed model has a more pronounced impact on the estimation of odds ratio than adjusting for selection bias. This considerable difference can be explained by the strong clustering effect present in these data. Nevertheless, there is a greater difference in the *p*-values of testing the null hypothesis between using bias reduction methods or not than adjusting for dependent data structures.

Although the different analysis methods present different results, they at least point in the same direction, indicating that the estimated probability of receiving a case conference is higher in the treatment group than in the control group. However, in our study, when adjustment for bias and dependencies is performed, the null hypothesis of a difference in the use of condition between the two groups could not be rejected. Although there is a hint that there could be a difference, this difference could not be detected in this study due to the resulting sample and the limits of the study design.

Table 4
Estimated probabilities and differences between groups, odds and odds ratios using different models – GLM = generalised linear model or GLMM = generalised linear mixed model, and GLMM with additional methods for bias reduction, whereby 1 = genetic matching using propensity score, 2 = genetic matching using covariates, 3 = covariate adjustment using propensity score, 4 = covariate adjustment using several covariates. The table shows the marginal treatment effect, the odds ratio for both groups and the risk difference. The conditional treatment effects given by the model are also shown in brackets. The *p*-value is the probability of the Wald test statistic for the null hypothesis of no difference between the two groups.

	$p_{Control}$	p_{Treat}	Difference	$Odd_{Control}$	Odd_{Treat}	Odds ratio	*p*
GLM	0.80	0.91	0.11	3.95	10.18	2.58	< 0.01
GLMM	0.86	0.98	0.12	6.28	51.72	8.23	0.03
1	0.91	0.98	0.07	9.72	46.78	4.81	0.14
2	0.91	0.98	0.06	10.35	40.33	3.90	0.16
3	0.79 (0.83)	0.92 (0.97)	0.12 (0.14)	3.84 (4.98)	11.18 (31.40)	2.91 (6.30)	0.07
4	0.80 (0.73)	0.91 (0.95)	0.12 (0.22)	3.90 (2.72)	10.63 (18.98)	2.72 (6.99)	0.10

4.2. Model choice for estimation

In Fig. 1, the model choice is determined by the data characteristics, e.g., outcome type and study type and design. Our example shows that choosing the model to estimate effects within observational studies is also closely related to different key issues, such as unobserved variables, sample size and the study objectives. Here, we adjust for clustered data by using the generalised linear mixed model, where several methods are available to reduce selection bias: genetic matching and regression with propensity score or covariates. In our opinion, no single method for bias adjustment is optimal, and each approach has its own limitations and is open for discussion.

Propensity score. Since Rosenbaum and Rubin (1983), the propensity score has become increasingly popular for adjusting selection bias via matching methods or regression. However, there is a debate regarding the use of propensity scores to recover causal effects from observational studies. First, regression adjustment is not a recommended way to use the propensity score (Austin et al., 2007). Second, the propensity score is criticized for having the drawback of losing potentially useful information about predictors of outcomes (Stürmer et al., 2006). However, using the propensity score as the matching variable can circumvent the problem of having too many variables (Cepeda et al., 2003). Additionally, it has the advantage of balancing on a large number of covariates in one summarising variable, where finding matches for a large number of variables is nearly impossible (Starks et al., 2009).

Genetic matching. The genetic matching approach provides the ability to balance group allocation to imitate randomisation, such as the bias adjustment being independent from the outcome. Because only a limited number of covariates for adjustment could be used, traditional matching is often limited (D'Agostino, 1998). Hence, instead of the propensity score being used as the summary, it can be used to balance the covariates in two groups. However, regardless of whether the propensity score or covariates is employed, using the matching approaches has the trade-off of losing a large proportion of observations, which then influences the estimand (Wang, 2009). In our example, this resulted in 246 observations, some of which are sampled more than once to receive the matched sample. Furthermore, the resulting samples could be prone to hidden bias in unobserved variables, which is a general problem for all observational methods. Hence, in our example, there is a strong assumption of hidden bias after applying the methods. Ridder and Graeve (2011) stated that if hidden bias is present, then matching using the propensity score has a comparable bias, but the precision of the estimates is lower. However, matching provides balance checks and hidden bias assessment, which is not possible in a regression framework. Furthermore, after this approach, the sample can be analysed with small and simple models to estimate the interested effects. In our examples, this model has only one fixed effect and a random structure.

Regression. Bias adjustment via regression includes the adjustment variables within the estimation model; thus, the additional subsequent step of sample matching is not required. Since the simulation study of Wang (2009) found that using the propensity score provides biased effect estimates, the advantage of using the propensity score within regression to adjust for known confounders was demonstrated in small datasets by Biondi-Zoccai et al. (2011) and particularly for dichotomous outcomes (Cepeda et al., 2003). The reason for this result is that adding more variables can decrease statistical power in small samples and that using the propensity score instead produces similar estimation results with limited power (Starks et al., 2009). Using all covariates as fixed effects in the model instead of the summarised value of propensity score is also possible. Then, no information is lost, but unfortunately, the convergence of full models is often not possible or the confidence intervals of the estimates are too large because of the small number of observations in each case. This was the case in our study example, where only a subset of all covariates could be included in the regression model.

4.3. Study design and sample size in observational studies

On the one hand, in the literature, there is a demand for robustly designed observational studies to avoid as much selection bias as possible (Ellenberg, 1994), for example, a high participation rate to achieve a representative sample of the population (Hammer et al., 2009). On the other hand, if selection bias occurs and one adjustment method has to be chosen, the goal is to obtain the best method for removing bias while ensuring optimal estimation results. Therefore, a very precise estimate is not useful if it is drastically wrong, and thus, an estimate with a small bias rather than a small variance should be more convincing (Rubin, 2006). Hence, before collecting data for an observational study, two major concerns should be taken into account: (1) covariates that may obtain selection bias and hence require measurement are determined and (2) a larger sample size is needed to ensure a sufficient sample size; although there is a loss due to adjustment methods, the former concern has the most important effect. There is a demand for the ability to calculate the sample size that is needed for a sufficient estimate quality, although there is also a need to adjust for an assumed selection bias before the data are collected. Hence, further investigations should be performed to permit drawing conclusions regarding the minimum required sample size within observational studies, which has to be adjusted for bias, or, if bias appears, how much of the sample is being lost via matching.

4.4. Limitations

Several limitations of our presented model should be discussed and explained.

First, although the participants were nested in care units, which in turn were nested in nursing homes, we decided to use a 2-level instead of a 3-level mixed model for analysis, whereby the care units could also explain a part of the variation (for intra-class correlation coefficient estimates of different levels from the regression model, see the Vignette). Due to the given data structure – more than half of the nursing homes provided participants in only one care unit – we chose a simple model only using nursing homes as random effects, knowing that more than half of the variability was explained by it.

Second, we could not provide the balance of various characteristics at the different levels of the data, i.e., not at the care unit level (Leyrat et al., 2014) or at the nursing home level (Belitser et al., 2011). Due to our decision to use nursing homes as a level of clustering, an adjustment at that level has some limits. In our example, we ignore the multilevel structure in both the propensity score estimation model and the matching implementation. On the one hand, the most straightforward idea is to force matching within each cluster, provided that treated and control cases are available within each cluster. However, this approach is very difficult to realise when the cluster sizes are small and may yield a considerable loss of individuals (Arpino and Cannas, 2016). On the other hand, Arpino and Mealli (2011) proposed including the level of clustering in the propensity score estimation model if the treatment is assigned at the individual level in multilevel settings. Then, the adjustment via matching is not forced within clusters. However, in our example, this model for the propensity score estimation did not converge, and we could provide bias adjustment at only the individual level. In this context, Li et al. (2013) show that accounting for a cluster structure in at least one stage, e.g., in the propensity score estimation or within the outcome model, can greatly reduce the bias. Nevertheless, upcoming studies faced with both existing bias and multilevel data should consider applying the modern methods of optimal multilevel matching (Pimentel et al., 2015; Zubizarreta and Keele, 2014), which can close the gap.

Furthermore, we did not use any procedure for model selection (e.g., the iterative process of the estimation and balance check) for the propensity score estimation itself. This decision was pragmatic and was based on using the individual related characteristics, which were

obtained in this observational study and assigned in the literature as being correlated with the circumstance of being a resident of dementia specific care units or not. Nevertheless, studies with more possible given covariates and a larger sample size and with the aim of using propensity scores primarily for bias reduction should be considered during model selection in combination with balance checks (Belitser et al., 2011).

We also did not consider the lack of independence assumption, either for the regression model after matching or for assessing hidden bias using Rosenbaum bounds, due to the matching with replacement. Solutions for this induced problem that use matching methods were reported by Stuart (2010) and Zubizarreta and Keele (2014).

Finally, for handling the missing data of the 53 participants, a statistical strategy such as multiple imputation was not conducted, nor were the results validated using a sensitivity analysis (for more details, see Carpenter and Kenward, 2013; Little and Rubin, 2002). Such a complete case analysis reduces statistical power and estimate precision; additionally, estimates can be biased in some circumstances if the missing data are not randomly distributed (Bartlett et al., 2015).

Acknowledgements

The data used for the analysis in this paper were collected in the study DemenzMonitor (2012–2014), which was funded by the DZNE Site Witten. The principal investigator of the DemenzMonitor Study is Dr. Bernhard Holle.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijnurstu.2017.06.017.

References

American Association of Colleges of Nursing, 2015. Nursing Research – Position Statement. http://www.aacn.nche.edu/publications/position/nursing-research.
Arpino, B., Cannas, M., 2016. Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. Stat. Med. 35 (12), 2074–2091.
Arpino, B., Mealli, F., 2011. The specification of the propensity score in multilevel observational studies. Comput. Stat. Data Anal. 55 (4), 1770–1780.
Austin, P.C., 2007. The performance of different propensity score methods for estimating marginal odds ratios. Stat. Med. 26 (16), 3078–3094.
Austin, P.C., 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. 28 (25), 3083–3107.
Austin, P.C., Grootendorst, P., Anderson, G.M., 2007. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat. Med. 26 (4), 734–753.
Bartlett, J.W., Harel, O., Carpenter, J.R., 2015. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. Am. J. Epidemiol. 182 (8), 730–736.
Baser, O., 2006. Too much ado about propensity score models? Comparing methods of propensity score matching. Value Health 9 (6), 377–385.
Belitser, S.V., Martens, E.P., Pestman, W.R., Groenwold, R.H., de Boer, A., Klungel, O.H., 2011. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol. Drug Saf. 20 (11), 1115–1129.
Biondi-Zoccai, G., Romagnoli, E., Agostoni, P., Capodanno, D., Castagno, D., D'Ascenzo, F., Sangiorgi, G., Modena, M.G., 2011. Are propensity scores really superior to standard multivariable analysis? Contemp. Clin. Trials 32 (5), 731–740.
Burns, N., Grove, S.K., 2009. The Practice of Nursing Research. Appraisal, Synthesis and Generation of Evidence, 6th ed. Saunders Elsevier, St. Louis.
Carpenter, J.R., Kenward, M.G., 2013. Multiple Imputation and Its Application. John Wiley & Sons Ltd, Chichester, UK.
Cepeda, M.S., Boston, R., Farrar, J.T., Strom, B.L., 2003. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am. J. Epidemiol. 158 (3), 280–287.
D'Agostino, R.B., 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat. Med. 17 (19), 2265–2281.
Ellenberg, J.H., 1994. Selection bias in observational and experimental studies. Stat. Med. 13 (5–7), 557–567.
Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research. New York, Cambridge University Press.
Greenland, S., 1987. Interpretation and choice of effect measures in epidemiologic analyses. Am. J. Epidemiol. 125, 761–768.
Hammer, G.P., du Prel, J.B., Blettner, M., 2009. Avoiding bias in observational studies: Part 8 in a series of articles on evaluation of scientific publications. Deutsches Ärzteblatt Int. 106 (41), 664–668.
Hardin, J., Hilbe, J., 2012. Generalized Estimating Equations, 2nd ed. CRC Press, Boca Raton, FL.
Leyrat, C., Caille, A., Donner, A., Giraudeau, B., 2014. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. Stat. Med. 33 (20), 3556–3575.
Li, B., Lingsma, H.F., Steyerberg, E.W., Lesaffre, E., 2011. Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. BMC Med. Res. Methodol. 11 (1), 77.
Li, F., Zaslavsky, A.M., Landrum, M.B., 2013. Propensity score weighting with multilevel data. Stat. Med. 32 (19), 3373–3387.
Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis With Missing Data. John Wiley & Sons, Inc., Hoboken, NJ.
Nguyen, T.-L., Collins, G.S., Spence, J., Daurès, J.-P., Devereaux, P.J., Landais, P., Le Manach, Y., 2017. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. BMC Med. Res. Methodol. 17 (1), 78.
O'Connor, A., 2013. Interpretation of odds and risk ratios. J. Vet. Intern. Med. 27 (3), 600–603.
Ostir, G.V., Uchida, T., 2000. Logistic regression: a nontechnical review. Am. J. Phys. Med. Rehabil. 79 (6), 565–572.
Palm, R., Bartholomeyczik, S., Roes, M., Holle, B., 2014. Structural characteristics of specialised living units for people with dementia: a cross-sectional study in German nursing homes. Int. J. Ment. Health Syst. 8.
Palm, R., Trutschel, D., Simon, M., Bartholomeyczik, S., Holle, B., 2015. Differences in case conferences in dementia specific vs traditional care units in German nursing homes: results from a cross-sectional study. J. Am. Med. Dir. Assoc. 17 (1), 91.e9–91.e13.
Pimentel, S.D., Yoon, F., Keele, L., 2015. Variable-ratio matching with fine balance in a study of the peer health exchange. Stat. Med. 34 (30), 4070–4082.
R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Randolph, J.J., Falbe, K., Manuel, A.K., Balloun, J.L., 2014. A step by step guide to propensity score matching in R. Pract. Assess. Res. Eval. 19 (18).
Ridder, A.D., Graeve, D.D., 2011. Can we account for selection bias? A comparison between bare metal and drug-eluting stents. Value Health 14 (1), 3–14.
Rosenbaum, P., 2002. Observational Studies. Springer Series in Statistics. Springer, New York.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.
Rubin, D.B., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge, UK.
Sekhon, J.S., 2011. Multivariate and propensity score matching software with automated balance optimization: the Matching package for R. J. Stat. Softw. 42 (7), 1–52.
Starks, H., Diehr, P., Curtis, J.R., 2009. The challenge of selection bias and confounding in palliative care research. J. Palliat. Med. 12 (2), 181–187.
Stroup, W., 2012. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, Boca Raton, FL.
Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Stat. Sci. Rev. J. Inst. Math. Stat. 25 (1), 1–21.
Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., Schneeweiss, S., 2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J. Clin. Epidemiol. 59 (5), 437.e1–437.e24.
Wang, Z., 2009. Propensity score methods to adjust for confounding in assessing treatment effects: bias and precision. Internet J. Epidemiol. 7 (2).
Williamson, E., Morley, R., Lucas, A., Carpenter, J., 2012. Propensity scores: from naive enthusiasm to intuitive understanding. Stat. Methods Med. Res. 21 (3), 273–293.
Wu, S., Crespi, C.M., Wong, W.K., 2012. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp. Clin. Trials 33 (5), 869–880.
Zou, G., Donner, A., 2004. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 60 (3), 807–811.
Zubizarreta, J.R., Keele, L., 2014. Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System. arXiv preprint arXiv:1409.8597.
Zúñiga, F., Ausserhofer, D., Hamers, J.P., Engberg, S., Simon, M., Schwendimann, R., 2015. Are staffing, work environment, work stressors, and rationing of care related to care workers' perception of quality of care? A cross-sectional study. J. Am. Med. Dir. Assoc. 16 (10), 860–866.