Open Journal of Epidemiology, 2015, 5, 155-162
Published Online August 2015 in SciRes. http://www.scirp.org/journal/ojepi
http://dx.doi.org/10.4236/ojepi.2015.53020

Choosing a Method to Reduce Selection Bias: A Tool for Researchers

Claire Keeble[1], Graham Richard Law[1], Stuart Barber[2], Paul D. Baxter[1]

[1]Division of Epidemiology and Biostatistics, University of Leeds, Leeds, UK
[2]Department of Statistics, University of Leeds, Leeds, UK
Email: c.m.keeble@leeds.ac.uk

Received 8 May 2015; accepted 6 July 2015; published 9 July 2015

Abstract

Selection bias is well known to affect surveys and epidemiological studies. There have been numerous methods proposed to reduce its effects, so many that researchers may be unclear which method is most suitable for their study; the wide choice may even deter some researchers, for fear of choosing a sub-optimal approach. We propose a straightforward tool to inform researchers of the most promising methods available to reduce selection bias and to assist the search for an appropriate method given their study design and details. We demonstrate the tool using three examples where selection bias may occur; the tool quickly eliminates inappropriate methods and guides the researcher towards those to consider implementing. If more studies consider selection bias and adopt methods to reduce it, valuable time and resources will be saved, and should lead to more focused research towards disease prevention or cure.

Keywords

Selection Bias, Participation Bias, Non-Response, Methodology

1. Introduction

Selection bias is known to affect health surveys and epidemiological studies [1] and can cause results from different studies on the same area of research to disagree or conclude contradictory findings [2]. However even in recent studies, selection bias is still sometimes ignored or dismissed [3]. If selection bias can be reduced across studies and lead to more consistent findings, time and resources could be saved since fewer repeated studies would be required. These savings could be used to develop more focused areas of research, contributing to increased knowledge.

With many selection bias reducing methods to choose from, researchers may find the process daunting and

possibly be deterred from implementing a method, for fear of choosing an unsuitable approach. The implementation of a method to reduce selection bias may also be viewed by researchers as an undesirable feature of their study, which could lead to criticisms of their study design and data collection, or potentially reduced chances of publication. There may also be time constraints, whereby study results need to be presented or published within a specific time frame, or a funded piece of research completed. In these instances, research into selection bias and methods to reduce it may not be prioritized. This work aims to provide reassurance to researchers that applying a method to reduce selection bias is a positive aspect of their study which should be encouraged whenever selection bias is suspected. Consideration of selection bias as a possibility should be routine. An appropriate method should be applied and used as either a sensitivity analysis to reassure readers of the study results, or to produce findings with reduced bias. We aim to draw attention to the available methods to reduce selection bias and provide sources for further reading. We intend to give guidance for selecting a method, structured in such a way that it is applicable for any study or survey potentially affected by selection bias.

Various methods, which will be discussed here, have been suggested to reduce selection bias, each with their own requirements and assumptions. Some methods require there to be additional data available external to the study [4]-[7], some require data regarding the non-participants [8], and some assume that the variable associated with selection is known and measured [9]-[11]. Each method will be briefly introduced and sources given for further reading. A tool in the form of a straightforward flowchart is also provided to aid the selection of an appropriate method, depending upon the details of a study and any additional information available. Three examples are presented which demonstrate the flowchart. Exploration into the suggested methods can then be conducted, allowing the researcher to select a method to reduce selection bias more easily. We hope this tool encourages the use of such methods and consequently leads to results less affected by selection bias.

2. Methods to Reduce Selection Bias

Table 1 summarizes the main methods used in the literature to reduce selection bias [3] which will be included in this guidance, and gives suggestions for further reading through original articles, examples or comprehensive summaries. These methods have similar themes, such as using the variable associated with selection in the analysis, weighting responses or predicting the effects of the bias. Additional, less frequently used methods are of course applicable, as are any new methods which are not yet widely used, but this tool is designed to be a starting point which can be developed through time and which should be useful for most types of research affected

Table 1. Current key methods used to reduce selection bias.

Method	Brief description
Adjust for Selection Bias [9] [12]	Include the variable associated with selection in the analysis, to reduce selection bias in a similar way to confounding [13].
Bias Breaking [5]	A method which produces bias-adjusted estimates for the odds ratios in case control studies.
Imputation [8]	Often multiple imputation; replace missing values with reasonable estimates using the collected data.
Population Data [7]	Use population data in place of control data in a case control study.
Post-Stratification [14]	Classify unmatched samples of cases and controls based on their values on one or more of the variables in the study. Similar to stratified sampling or frequency matching.
Predict the Bias [15]-[17]	Use information from non-participants to try to predict the amount of bias present.
Propensity Score [10]	Can be used to match cases and controls, or as an additional covariate during analysis.
Sensitivity Analysis [6] [12]	A method for estimating the direction and magnitude of the bias.
Stratification [11]	Calculate estimates conditional on at least one other variable, which can lead to unbiased estimates within strata.
Weighting [4]	Usually inverse probability weighting; use external data to assign each subject a weight which is the inverse of their probability of selection, to allow them to represent non-participants.

by selection bias.

Table 2 includes the data requirements of each method; including whether it requires the variable associated with selection to be known and recorded, whether population data external to the study are required or whether data regarding the non-participants who declined the study are needed. Where more than one category has been ticked, this indicates the method can be adapted for use when an alternative source for the required data is available. For the variable associated with selection to be collected, it is assumed the variable is known and can be recorded during the study. There may be instances where this variable is unknown, cannot be collected, or is impractical to record. This includes data which are expensive to collect, sensitive, or due to an unidentified variable. However there may be instances where a proxy may be used instead. The population data indicates information external to the study, for example from a database or alternative records. Sources may include census data or hospital registries. These data are assumed to be unbiased and represent the entire population of interest. Non-participant data are basic characteristics recorded from those who were unwilling to participate. These are usually data from the subject themselves, but may also be from external sources similar to those used to collect the population data.

Although the three data categories used in **Table 2** are sourced from different places (the participants, the population and non-participants respectively), there are relationships between them. For example, if the original potential participants are representative of the population of interest, and relevant information is known for all non-participants, then the non-participant data in conjunction with the participant data could be used to approximate the population data. Therefore under certain circumstances it may be possible to use a different column from **Table 2** for the data source, other than the one(s) ticked. **Table 2** and the consequent tool can be interpreted as a generalization or guide, which can be adapted by the researcher if these conditions are met.

Although each of the methods in **Table 1** is designed to reduce selection bias, they do so using different techniques and assumptions. Therefore, a method which may be optimal for one study may not be suitable for another. Some are also aimed at particular study designs, for example two were developed specifically for case control studies [5] [7].

Several of the methods in **Table 1** have been developed or derived from one another. For example, the bias-breaking method [5] is a form of post-stratification, which is a type of stratification, and the propensity score is derived from stratification. However, their suitability as a method to reduce selection bias differs between studies. There are also similarities amongst some of the methods. For example, predicting the amount of bias present is similar to a sensitivity analysis, and several of the methods also began in survey literature [4]. **Figure 1** gives an example of a flowchart based on **Table 2** which could be used by researchers to shortlist potential selection bias reducing methods for further investigation. Researchers could extend this flowchart to meet their specific needs for the variables or datasets they encounter, or alternatively disciplines could form a subject-specific chart to which new methods could be added over time.

Table 2. The required data for the methods summarised in **Table 1**.

	Selection variable	Population data	Non-participant data
Adjust for Selection Bias	✓		
Bias Breaking		✓	
Imputation			✓
Population Data		✓	
Post-Stratification	✓		
Predict the Bias	✓	✓	✓
Propensity Score	✓		
Sensitivity Analysis	✓	✓	
Stratification	✓		
Weighting		✓	✓

3. Examples

Three examples of hypothetical studies follow which utilize the flowchart (**Figure 1**) to determine a suitable method to apply. The flowchart begins in the top-left corner, shown using a bold outline. To answer the first question in the flowchart, the requirements for selection bias to occur must be known. For selection bias to be present, there must be the exposure and outcome of interest, and these must both affect whether or not an individual is selected to participate or self-selects (participates) in the study. This selection variable must then be conditioned on, which it often will be since only those who have participated can be studied [1].

Once the flowchart has provided a list of possible methods to explore further, it is the responsibility of the researcher to consider each in turn to see which method is most suitable for their particular study. All method assumptions must be considered and the details of the specific study incorporated.

3.1. Example 1

A randomized controlled trial (RCT) is conducted for a new hayfever tablet. Hayfever suffers are recruited and randomly allocated to either the drug group or the placebo group. The new tablet produces some unexpected side-effects and some participants in the drug arm suffer from fainting or severe vomiting. Half of the participants in the drug arm withdraw from the study, as they decide that their hayfever symptoms are preferable to the side effects. The flowchart can be used to see which methods for selection bias may be worth further consideration.

- Is the study potentially affected by selection bias? The association of interest is from the new tablet, or treatment group, to the severity of the hayfever symptoms. For potential selection bias, both the treatment group and the hayfever symptoms need to influence selection into the study. The side-effects from the tablet causing withdrawal from the study mean that the treatment group does affect inclusion in the analysis and hence 'selection'. However, hayfever suffers were randomly allocated to either the drug or placebo group, so the severity of hayfever symptoms was balanced between the two treatment groups and therefore the severity of the symptoms did not affect selection into the study. Since only the treatment group and not the severity of the symptoms affects selection, selection bias is not a problem here, and the results can be analyzed as usual without the need for a selection bias reducing method.

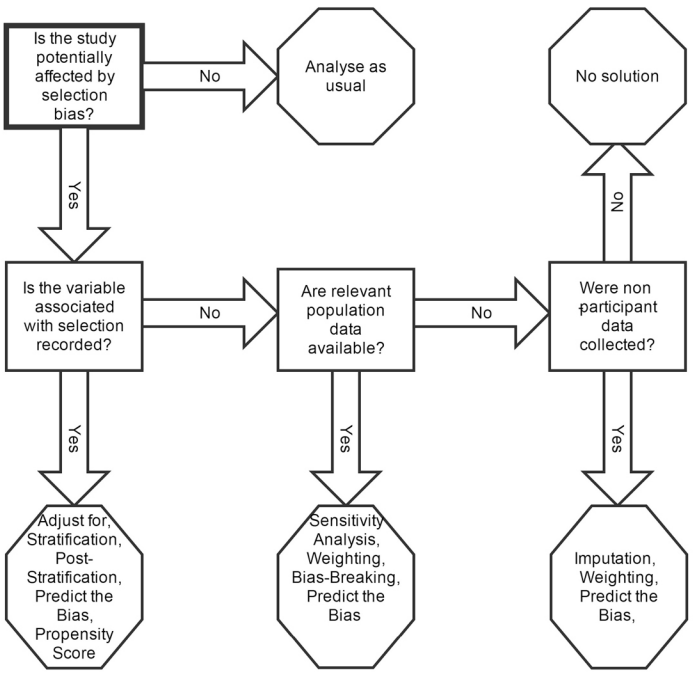

Figure 1. A flowchart tool for researchers: Which methods are suitable to reduce selection bias?

3.2. Example 2

A study is conducted investigating the association between coffee and migraines. Postal surveys are randomly sent to households in the United Kingdom (UK), with a return envelope enclosed. In addition to questions about migraines and coffee consumption, the survey also includes basic demographic data such as sex, age, general location, employment status etc. Those who drink coffee may be more likely to respond, to find out whether they are increasing their chances of migraines, whereas people who do not drink coffee may not be as interested. Those who suffer from migraines may also be more interested in the survey than those who do not. Previous studies have shown that older people are generally more likely to participate in surveys [18]-[22]. Let age and coffee consumption be positively correlated and let migraines be more common in older people. The flowchart can guide the researchers involved in the study towards any possible methods to reduce selection bias.

- Is the study potentially affected by selection bias? Coffee is the exposure of interest and migraines are the outcome of interest. Since coffee drinkers and migraine sufferers are more likely to return the survey, then selection is affected by both the exposure and outcome of interest, and therefore selection bias is possible.
- Is the variable associated with selection recorded? Coffee consumption, migraine occurrences and age all affect survey returns, and are all recorded in the survey. However, only variables which are not the exposure or outcome of interest can be used in the analysis to reduce selection bias; age in this instance. Therefore, the following methods can be considered further:

- Adjust for the variable associated with selection; Age can be added to the analysis, for example as a variable in a regression model between coffee and migraines.
- Stratification; The analysis can be conducted within age strata; for example by analyzing in age groups of ten years, to reduce the effect of age on selection.
- Post-stratification; Migraine sufferers can be matched to non-sufferers by their age.
- Predict the bias; The analysis could be conducted with and without the age variable, to predict how selection bias is affected by age.
- Propensity score; Age can be incorporated into the analysis to either match migraine sufferers to non-sufferers, or the propensity score can be calculated using all the variables and included in the analysis.

- Therefore all methods here are possible.

3.3. Example 3

A UK case control study is conducted which investigates the association between excessive alcohol consumption and brain tumors. Researchers therefore attempt to recruit cases who have brain tumors and controls who do not. The retrospective nature of the study design mean data are then collected on each participant regarding their alcohol consumption, in this instance over an extended previous period of time. The exposure of interest in the case control study is blinded to the participants and the interviewers used, to reduce the effects of other biases such as interviewer bias. The blinding here is in the form of an extended questionnaire, including questions regarding several possible exposures such as alcohol, smoking, mobile phone use, exercise routines and family history. This leads to some participants intentionally avoiding questions, such as those to which they have undesirable answers. For example, heavy smokers ignore the question regarding the number of cigarettes smoked daily, those who do not exercise often miss the question regarding the number of hours exercise completed per week, and frequent drinkers avoid the question about alcohol consumption.

Let the data from the questionnaire be available, along with a national database regarding the number of people with brain tumors in the UK. The Office for National Statistics (ONS) also records data for adult drinking habits [23]. The flowchart can be used to determine which methods may be suitable to reduce selection bias.

- Is the study potentially affected by selection bias? It is well-documented that cases are often more likely to participate in a study than controls, since they have additional motivation to find a cure or an explanation for their condition [24] [25]. Therefore the outcome is affecting self-selection into the study. Next the exposure of interest, alcohol consumption, is being recorded only for those who are willing to declare their consumption levels; in this instance, those who consume amounts not deemed to be excessive. Therefore, inclusion in the study analysis depends upon the exposure level. Since only those who are willing to participate in the study and who answer the question regarding alcohol consumption are used to investigate the association between excessive alcohol consumption and brain tumors, participation is conditioned on and so selection bias is a possibility.

- Is the variable associated with selection recorded? The variables associated with selection are the exposure and outcome themselves, hence methods which use the variable associated with selection in the analysis to reduce selection bias are not suitable here.
- Are relevant population data available? The national database for brain tumors and ONS data for drinking habits are available. Therefore, the following methods can be considered further:

- Sensitivity analysis; The external population data could be used to estimate the magnitude and direction of bias, although unfortunately the drinking habits of those in the population with brain tumors is unknown. This method may be possible.
- Weighting; If there are no heavy-drinking participants who answer the question about alcohol consumption, then a weight cannot be applied to this category and the weighing method would be unsuitable.
- Bias-breaking; For this method, a variable must be identified which separates the exposure from the selection criteria, but unfortunately, the risk factor *is* the variable which is determining selection into the analysis and hence this method is not suitable.
- Predict the bias; Information from non-participants is not available as such, but could possibly be derived from the population data available in conjunction with the participant data. This method may be possible.
- Population data; The method requires there to be data regarding the population size, the number of cases and the number of exposed; all of which are recorded in the national database or in the ONS records. Therefore this method is a possible option.

These examples have shown how the flowchart can quickly eliminate potential methods and guide the researcher towards a subset of methods for further consideration.

4. Discussion

Selection bias can be problematic for surveys and a range of study designs [26], but particularly those which are retrospective such as case control studies [2], as seen in Example 3. Biased results can lead to incorrect findings and the unnecessary repetition of studies, wasting valuable time and resources which could instead be used to fund additional research into diseases or their cure.

Any form of bias can be viewed as a negative aspect of a study, but action should be taken to reduce as much bias as possible within the results of a study. This work aims to highlight the importance of bias reduction, specifically selection bias, and provide researchers with a summary of methods currently available to reduce selection bias, along with references for further reading.

A user-friendly flowchart tool has been provided, which can be adapted for particular research areas or depending upon the data resources available, to aid the selection of an appropriate method. The tool is not designed to identify one method to use, but instead guide the researcher to a subset of methods for further consideration. This could be viewed as a limitation, but contemplation of the requirements for each method is necessary. The optimal method depends upon specific details relating to an individual study and would require a complicated flowchart which would be more difficult to use. However, subject-specific flowcharts could be created. This tool is therefore a straightforward flowchart, applicable to a range of study designs, provoking consideration of selection bias while providing references for further reading. We hope demonstration of this versatile tool through examples and raised awareness results in more consideration of selection bias and consequently the implementation of appropriate methods.

5. Conclusion

Bias reduction is an important part of any study and this work raises awareness of selection bias in particular. A straightforward flowchart, with summaries of the current methods to reduce selection bias, has been provided to guide researchers towards a suitable method, in the hope that more accurate results are generated from studies.

Funding

Claire Keeble is funded by an MRC Capacity Building Studentship. Paul D Baxter, Stuart Barber and Graham Richard Law are funded by HEFCE. The funding sources had no involvement in the study design, in the collection, analysis and interpretation of data, in the writing of the report or the decision to submit the article for publication.

References

[1] Hernan, M., Hernandez-Diaz, S. and Robins, J. (2004) A Structural Approach to Selection Bias. *Epidemiology*, **15**, 615-625. http://dx.doi.org/10.1097/01.ede.0000135174.63482.43

[2] Hennekens, C.H. and Buring, J.E. (1987) Screening. In: Mayrent, S.L., Ed., *Epidemiology in Medicine*, Little, Brown and Co., Boston, 327-345.

[3] Keeble, C., Barber, S., Law, G. and Baxter, P. (2013) Participation Bias Assessment in Three High Impact Journals. *Sage Open*, **3**, 1-5. http://dx.doi.org/10.1177/2158244013511260

[4] Horvitz, D. and Thompson, D. (1952) A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685. http://dx.doi.org/10.1080/01621459.1952.10483446

[5] Geneletti, S., Richardson, S. and Best, N. (2009) Adjusting for Selection Bias in Retrospective, Case-Control Studies. *Biostatistics*, **10**, 17-31. http://dx.doi.org/10.1093/biostatistics/kxn010

[6] Geneletti, S., Mason, A. and Best, N. (2011) Adjusting for Selection Effects in Epidemiologic Studies: Why Sensitivity Analysis Is the Only "Solution". *Epidemiology*, **22**, 36-39. http://dx.doi.org/10.1097/EDE.0b013e3182003276

[7] Keeble, C., Barber, S., Baxter, P., Parslow, R. and Law, G. (2014) Reducing Participation Bias in Case-Control Studies: Type 1 Diabetes in Children and Stroke in Adults. *Open Journal of Epidemiology*, **4**, 129-134. http://dx.doi.org/10.4236/ojepi.2014.43018

[8] Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., *et al*. (2009) Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, **338**, 2393-2397. http://dx.doi.org/10.1136/bmj.b2393

[9] Breslow, N. and Day, N. (1980) Chapter 3: General Considerations for the Analysis of Case-Control Studies. In: Breslow, N.E. and Day, N.E., Eds., *Statistical Methods in Cancer Research*, IARC Scientific Publications, 84-119.

[10] Rosenbaum, P. and Rubin, D. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55. http://dx.doi.org/10.1093/biomet/70.1.41

[11] Sarndal, C.E. (1992) Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used. *Survey Methodology*, **18**, 241-252.

[12] Kleinbaum, D., Morgenstern, H. and Kupper, L. (1981) Selection Bias in Epidemiological Studies. *American Journal of Epidemiology*, **113**, 452-463.

[13] Hosmer Jr., D., Lemeshow, S. and Sturdivant, R. (2013) Applied Logistic Regression. John Wiley & Sons, Hoboken. http://dx.doi.org/10.1002/9781118548387

[14] Schlesselman, J. (1982) Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, New York.

[15] Hatch, E., Kleinerman, R., Linet, M., Tarone, R., Kaune, W., Auvinen, A., *et al*. (2000) Do Confounding or Selection Factors of Residential Wiring Codes and Magnetic Fields Distort Findings of Electromagnetic Field Studies? *Epidemiology*, **11**, 189-198. http://dx.doi.org/10.1097/00001648-200003000-00019

[16] Madigan, M., Troisi, R., Potischman, N., Brogan, D., Gammon, M., Malone, K., *et al*. (2000) Characteristics of Respondents and Non-Respondents from a Case-Control Study of Breast Cancer in Younger Women. *International Journal of Epidemiology*, **29**, 793-798. http://dx.doi.org/10.1093/ije/29.5.793

[17] Wrensch, M. (2000) Are Prior Head Injuries or Diagnostic X-Rays Associated with Glioma in Adults? The Effects of Control Selection Bias. *Neuroepidemiology*, **19**, 234-244. http://dx.doi.org/10.1159/000026261

[18] Hara, M., Higaki, Y., Imaizumi, T., Taguchi, N., Nakamura, K., Nanri, H., *et al*. (2010) Factors Influencing Participation Rate in a Baseline Survey of a Genetic Cohort in Japan. *Journal of Epidemiology*, **20**, 40-45. http://dx.doi.org/10.2188/jea.JE20090062

[19] Perez, D., Nie, J., Ardern, C., Radhu, N. and Ritvo, P. (2013) Impact of Participant Incentives and Direct and Snowball Sampling on Survey Response Rate in an Ethnically Diverse Community: Results from a Pilot Study of Physical Activity and the Built Environment. *Journal of Immigrant and Minority Health*, **15**, 207-214. http://dx.doi.org/10.1007/s10903-011-9525-y

[20] Koloski, N., Jones, M., Eslick, G. and Talley, N. (2013) Predictors of Response Rates to a Long Term Follow-Up Mail out Survey. *PLoS ONE*, **8**, e79179. http://dx.doi.org/10.1371/journal.pone.0079179

[21] McLean, S., Paxton, S., Massey, R., Mond, J., Rodgers, B. and Hay, P. (2014) Prenotification but Not Envelope Teaser Increased Response Rates in a Bulimia Nervosa Mental Health Literacy Survey: A Randomized Controlled Trial. *Journal of Clinical Epidemiology*, **67**, 870-876. http://dx.doi.org/10.1016/j.jclinepi.2013.10.013

[22] Nota, S., Strooker, J. and Ring, D. (2014) Differences in Response Rates between Mail, E-Mail, and Telephone Follow-Up in Hand Surgery Research. *Hand*, **9**, 504-510. http://dx.doi.org/10.1007/s11552-014-9618-x

[23] Office for National Statistics (2013) Adult Drinking Habits in Great Britain.

http://www.ons.gov.uk/ons/rel/ghs/opinions-and-lifestyle-survey/adult-drinking-habits-in-great-britain–2013/stb-drinking-2013.html

[24] Galea, S. and Tracy, M. (2007) Participation Rates in Epidemiologic Studies. *Annals of Epidemiology*, **17**, 643-653. http://dx.doi.org/10.1016/j.annepidem.2007.03.013

[25] Li, Y., Wang, W., Wu, Q., van Velthoven, M., Chen, L., Du, X., *et al.* (2015) Increasing the Response Rate of Text Messaging Data Collection: A Delayed Randomized Controlled Trial. *Journal of the American Medical Informatics Association*, **22**, 51-64.

[26] Thadhani, R. and Tonelli, M. (2006) Cohort Studies: Marching Forward. *Clinical Journal of the American Society of Nephrology*, **1**, 1117-1123. http://dx.doi.org/10.2215/CJN.00080106