Open Journal of Epidemiology, 2014, 4, 129-134
Published Online August 2014 in SciRes. http://www.scirp.org/journal/ojepi
http://dx.doi.org/10.4236/ojepi.2014.43018

Reducing Participation Bias in Case-Control Studies: Type 1 Diabetes in Children and Stroke in Adults

Claire Keeble[1], Stuart Barber[2], Paul David Baxter[1], Roger Charles Parslow[1], Graham Richard Law[1]

[1]Division of Epidemiology & Biostatistics, University of Leeds, Leeds, UK
[2]Department of Statistics, University of Leeds, Leeds, UK
Email: mm07cmk@leeds.ac.uk

Received 29 May 2014; revised 29 June 2014; accepted 29 July 2014

Abstract

Background: Case-control studies have been used extensively in determining the aetiology of rare diseases. However, case-control studies often suffer from participation bias in the control group, resulting in biased odds ratios that cause problems with interpretation. Participation bias can be hard to detect and is often ignored. Methods: Population data can be used in place of the possibly biased control group, to investigate whether participation bias may have affected the results in previous studies, or in place of controls in future studies. We demonstrate this approach by reanalysing and comparing the results of two case-control studies: Type 1 diabetes in Yorkshire children and stroke in Indian adults. Findings: Using population data to represent the control groups reduced the width of the confidence intervals given in the original studies and confirmed the findings for the two diabetes risk factors used; caesarean birth (odds ratio (OR) = 2.12 (1.53, 2.95) compared with 1.84 (1.09, 3.10)) and amniocentesis (OR = 3.38 (2.09, 5.47) compared with 3.85 (1.34, 11.04)). The three stroke risk factors investigated were found to have increased odds ratios when using population data; hypertension (OR = 5.645 (5.639, 5.650) compared with 3.807 (2.114, 6.856)), diabetes (OR = 12.212 (12.200, 12.224) compared with 3.473 (1.757, 6.866)) and smoking (OR = 5.701 (5.696, 5.707) compared with 2.242 (1.255, 4.005)). Interpretation: Participation bias can greatly affect the results of a study and cause some potential risk factors to be over- or underestimated. This approach allows previous studies to be investigated for participation bias and presents an alternative to a control group in future studies, while improving precision.

Keywords

Case-Control, Diabetes, Participation Bias, Stroke, Selection Bias

How to cite this paper: Keeble, C., Barber, S., Baxter, P.D., Parslow, R.C. and Law, G.R. (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

1. Introduction

Participation bias, a subset of selection bias, affects many study types and is often ignored by authors [1]. It is well documented that case-control studies can be affected by participation bias in the control group [2]-[4], which can result in an over- or underestimation of odds ratios [5].

In recent years, routine data has become more widely available; partially due to advances in technology, increased routine data collection and emphasis on data sharing, along with the recent move towards and focus on Big Data. Linked data sources such as hospital episode statistics (HES) [6], the clinical practice research database (CPRD) [7] and Research One [8] are allowing information to be shared more easily and further research to be carried out. Often these databases hold much more information, on a greater number of people, than could easily be collected through a study. Some census databases also contain information relating to every member in a population [9] [10].

We propose the use of population data in place of control data, along with the case data from a case-control study. We demonstrate this method by reanalysing a Yorkshire childhood diabetes case-control study and an Indian study of stroke. We explain how potential participation bias can be identified and show how to improve precision of the estimated odds ratios. We therefore present a method to reduce the amount of bias from the control group; which can be used in place of controls in future case-control studies to save time and resources, or as an approach to evaluate the results from previous studies.

2. Methods

2.1. The Data

The diabetes data set used was taken from a case-control study [11], which had recorded cases of children under 16 years diagnosed with insulin-dependent diabetes mellitus (IDDM), or Type 1 diabetes, while resident in the area of the former Yorkshire Regional Health Authority, since 1978, with data collected 1993-1994. The stroke data set used 100 computed tomography (CT) proved cases of stroke, with age and sex matched controls, from hospital attendees in India [12]. These data sets have been used to demonstrate the effect of participation bias on the analysis of risk factors, and the potential for population data to provide improved estimates. The published results have been compared with results generated when population data is used in place of control data.

2.2. The Population Data

There are three values required from the population for each odds ratio replicated, which must be correct for the time and location of the original study:

1) The exposure in the population;
2) The size of the population;
3) The number of cases in the population.

For these examples, various sources were used, but all were publicly accessible to demonstrate the ease of the method (**Table 1**). However, more recent or detailed data could be obtained from previous studies or databases if available, which would be likely to improve the accuracy of the results.

2.3. The Proposed Method

The steps required to use population data in place of control data are as follows:

1) Use the population and case numbers to calculate the number of controls.
2) Use the exposed population and exposed case data to calculate the number of exposed controls.
3) Use the previous steps to calculate the remaining number of unexposed population, cases and controls.
4) Use these values to calculate odds ratios from a contingency table or using logistic regression.

These steps are shown below for the caesarean exposure in the diabetes data set as an example. This was repeated for exposures in both the diabetes and stroke data sets, using the methods used in the original study. The odds ratios published were also replicated, with all calculations using R [22].

Example, Caesarean:

$$\text{population} = \text{cases} + \text{controls}$$
$$774{,}840 = 248 + \text{controls}$$
$$774{,}840 = 248 + 774{,}592$$

Table 1. Population data used for the proposed method.

Diabetes data set			
Required population data	Specific requirement	Value collected	Source
1. Exposure in the population	Caesarean births in Yorkshire	9% of births	Birth Choice UK website [13]
	Amniocenteses in Yorkshire	15,000 in Britain each year	Cambridge Fetal Care [14]
2. Size of the population	Number of children in Yorkshire	774,840	Office of population censuses and surveys [15]
3. The number of cases in the population	Diabetes cases in Yorkshire	248	Yorkshire Childhood Diabetes Register [11]
Stroke data set			
Required population data	Specific requirement	Value collected	Source
1. Exposure in the population	Hypertension in India	23%	World Health Statistics [16]
	Diabetes in India	65.1 million	International Diabetes Federation [17]
	Smoking in India	14.925%	World Bank [18] [19]
2. Size of the population	Population size of India	1.237 billion	World Bank [20]
3. The number of cases in the population	Stroke cases in India	18,012,222	Rightdiagnosis.com [21]

$$\text{exposed population} = \text{exposed cases} + \text{exposed controls}$$

$$(0.09 \times 774{,}840) = \left(\frac{34}{196} \times 248\right) + \text{exposed controls}$$

$$69{,}736 = 43 + 69{,}693$$

$$\text{not exposed population} = \text{not exposed cases} + \text{not exposed controls}$$
$$(774{,}840 - 69{,}736) = (248 - 43) + (774{,}592 - 69{,}693)$$

$$705{,}104 = 205 + 704{,}899$$

This can be written generally; let *P* be the number of people in the population of interest, *D* be the disease of interest, *E* be the exposure of interest, *a* be the number of exposed cases and *c* be the number of unexposed cases. Values from the population can then be substituted into the equations below. The necessary steps are in bold.

$$P = P_{D=1} + P_{D=0}$$

$$\boldsymbol{P_{D=0} = P - P_{D=1}}$$

$$P_{E=1} = P_{D=1,E=1} + P_{D=0,E=1}$$

$$P_{D=1,E=1} = \frac{a}{a+c} \times P_{D=1}$$

$$\boldsymbol{P_{D=0,E=1} = P_{E=1} - P_{D=1,E=1}}$$

$$P_{E=0} = P_{D=1,E=0} + P_{D=0,E=0}$$

$$\boldsymbol{P_{E=0} = P - P_{E=1}}$$

$$\boldsymbol{P_{D=1,E=0} = P_{D=1} - P_{D=1,E=1}}$$

$$\boldsymbol{P_{D=0,E=0} = P_{D=0} - P_{D=0,E=1}}$$

3. Results

Table 2 shows the odds ratios and confidence intervals calculated using the population values, along with the

Table 2. Odds ratios and 95% confidence intervals comparing the published odds ratios with those generated using population data.

Data set	Exposure of interest	Published odds ratio (95% confidence interval)	Population data odds ratio (95% confidence interval)
Diabetes	Caesarean	1.81 (1.07, 3.04)	2.12 (1.53, 2.95)
	Amniocentesis	3.85 (1.34, 11.04)	3.38 (2.09, 5.47)
Stroke	Hypertension	3.807 (2.114, 6.856)	5.645 (5.639, 5.650)
	Diabetes	3.473 (1.757, 6.866)	12.212 (12.200, 12.224)
	Smoking	2.242 (1.255, 4.005)	5.701 (5.696, 5.707)

published odds ratios from the corresponding original study. It can be seen from **Table 2** that the population odds ratios support the findings from the original analysis of significantly raised odds ratios for birth by caesarean and amniocentesis in the diabetes data set. The results for the stroke data set all have increased odds ratios for the population data when compared with the initial study, however the confidence intervals of the hypertension population odds ratio and the published odds ratio do overlap. This could suggest support from the population data for the hypertension odds ratio but possible disagreement between the published and population odds ratios for the exposures diabetes and smoking; with greater disagreement when considering diabetes. One possible cause for this disagreement could be participation bias. Note the controls in the Indian stroke study were hospital attendees; this could have resulted in Berkson's bias [23], since those who smoke, have hypertension, or have diabetes, may have associated conditions requiring hospital admission. This higher proportion of smokers, hypertensive controls and diabetics in the control group than in the population would have resulted in lower odds ratios in the published results. Hence participation bias is likely to have occurred.

Table 2 also shows the population odds ratios have much narrower confidence intervals than the published odds ratios. This corresponds to the increase in the number of subjects considered in the population odds ratios compared with the number in the original case-control study.

4. Discussion

Participation bias can cause the results from studies to be inaccurate [5], especially in case-control studies where certain potential controls are more likely to participate than others. Researchers who may wish to use our method in place of, or in conjunction with, case-control studies, may have access to medical records or similar information which is likely to give more accurate odds ratios which are less affected by participation bias. In addition, the proposed method allows the identification of participation bias, as shown in the Indian stroke example, where Berkson's bias has been suggested. The method can also be extended to allow for matching in the original study, by stratifying or adding the confounder to the regression model, using more detailed population data, such as young and old stroke cases or male and female smokers.

Approximations may need to be made when data are available but not in the required format. For example, it was assumed that the number of 15 year old in Yorkshire was approximately a fifth of the 15 - 19 year old Yorkshire population [15]. Matched case-control studies may also be more time-consuming as more detailed population data are required, along with the confounding variable data for the cases. The case data will be available for new studies, but may not always be available for past studies. This was true for the Indian stroke study, where an unmatched analysis was required as an approximation, since the details linking the confounding variables to the cases were not published. As data availability has increased over the last few decades and census questions have become more detailed, similar population data for studies more recent than the diabetes study may be more readily available. It can, however, still be used as a tool to revisit older studies to confirm or question their findings. It is also likely that those working in these research areas would have access to databases or information from previous studies, allowing more accurate population data to be used. There will be circumstances where the required relevant population data will not be available and then a case-control study would be preferable.

This proposed method of using population data is very simple and quick to apply; far cheaper and easier than recruiting controls for a case-control study. This approach allows the study time and resources to be focused on the collection of case data, giving a larger sample of cases than previously possible. The method allows an efficient way to conduct a new large study, with less effort in the control group than previously required. The population data, if carefully selected, is likely to have reduced participation bias when compared with the corresponding control data, yielding more accurate results and increasing the chances of determining the true cause of a disease. Ideal sources of population data are those which capture information from the entire population of interest and which are considered to be reliable. Examples include population wide health databases or appropriate census data. However, if a population value is used and later thought to be inaccurate, the calculations can easily be rerun to generate improved estimates. The larger sample sizes resulting from this approach also generate narrower confidence intervals, allowing easier categorisation of the variables to significant protective factor, significant risk factor or insignificant risk or protective factor. All steps in the method were conducted using case information only in the paper, without the need for the original data set. Therefore, this analysis could be repeated for all variables published, to see whether any potential risk factors have been miscategorised. This method can support the findings from the study, or identify any potential bias in the results.

Identifying the true causes or risk factors of a disease is an important step towards developing a cure or preventing others from becoming cases. Case-control studies are a useful study design to help find the causes of a rare disease, but they can be affected by participation bias. A simple amendment to the method, such as the one proposed here, could help to yield more accurate results and move closer towards discovering the cause of the disease.

Funding

Claire Keeble is a Ph.D. student funded by a MRC Capacity Building Studentship. Graham Law, Stuart Barber, Paul Baxter and Roger Parslow are funded by HEFCE.

References

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

[2] Haapea, M., Miettunen, J., Veijola, J., Lauronen, E., Tanskanen, P. and Isohanni, M. (2007) Nonparticipation May Bias the Results of a Psychiatric Survey—An Analysis from the Survey Including Magnetic Resonance Imaging within the Northern Finland 1966 Birth Cohort. *Social Psychiatry and Psychiatric Epidemiology*, **42**, 403-409. http://dx.doi.org/10.1007/s00127-007-0178-z

[3] Lopez, R., Frydenberg, M. and Baelum, V. (2008) Non-Participation and Adjustment for Bias in Casecontrol Studies of Periodontitis. *European Journal of Oral Sciences*, **116**, 405-411. http://dx.doi.org/10.1111/j.1600-0722.2008.00567.x

[4] Tam, C.C., Higgins, C.D. and Rodrigues, L.C. (2011) Effect of Reminders on Mitigating Participation Bias in a Case-Control Study. *BMC Medical Research Methodology*, **11**, 33. http://dx.doi.org/10.1186/1471-2288-11-33

[5] Mezei, G. and Kheifets, L. (2006) Selection Bias and Its Implications for Case-Control Studies: A Case Study of Magnetic Field Exposure and Childhood Leukaemia. *International Journal of Epidemiology*, **35**, 397-406.

[6] Eckmann, C., Wasserman, M., Latif, F., Roberts, G. and Beriot-Mathiot, A. (2013) Increased Hospital Length of Stay Attributable to Clostridium Difficile Infection in Patients with Four Co-Morbidities: An Analysis of Hospital Episode Statistics in Four European Countries. *European Journal of Health Economics*, **14**, 835-846. http://dx.doi.org/10.1007/s10198-013-0498-8

[7] Childs, T., Scowcroft, A. and Todd, S. (2013) Gender and Regional Differences in the Treatment for Hypertension: A Pharmacoepidemiological Analysis of the General Practice Research Database (GPRD) in the Context of Hypertension in Atrial Fibrillation (AF) Patients. *Journal of Human Hypertension*, **27**, 648.

[8] Crossfield, S.S.R. and Clamp, S.E. (2013) Electronic Health Records Research in a Health Sector Environment with Multiple Provider Types. *HEALTHINF* 2013 *Proceedings of the International Conference on Health Informatics*.

[9] Sortsø, C., Thysegen, L.C. and Brønnum-Hansen, H. (2011) Database on Danish Population-Based Registers for Public Health and Welfare Research. *Scandinavian Journal of Public Health*, **39**, 17-19. http://dx.doi.org/10.1177/1403494811399171

[10] Ludvigsson, J.F., Otterblad-Olausson, P., Pettersson, B.U. and Ekbom, A. (2009) The Swedish Personal Identity Number: Possibilities and Pitfalls in Healthcare and Medical Research. *European Journal of Epidemiology*, **24**, 659-667. http://dx.doi.org/10.1007/s10654-009-9350-y

[11] McKinney, P.A., Parslow, R., Gurney, K., Law, G., Bodansky, H.J. and Williams, D.R.R. (1997) Antenatal Risk Fac-

tors for Childhood Diabetes Mellitus; A Case-Control Study of the Medical Record Data in Yorkshire, UK. *Diabetologia*, **40**, 933-939.

[12] Sorganvi, V., Kulkarni, M.S., Kadeli, D. and Atherga, S. (2014) Risk Factors for Stroke: A Case Control Study. *IJCRR*, **6**, 46-52.

[13] Birth Choice UK (2011) Graphs Of Historical Caesarean Section Rates. www.birthchoiceuk.com

[14] Cambridge Fetal Care (2013) Amniocentesis Test. www.fetalcare.co.uk

[15] Office of Population Censuses and Surveys (1995) Subnational Population Projections, Series PP3, No. 9, Table 5: 1993-Based Population Projections, 1993-2016: Sex and Quinary Age-Groups, p. 61.

[16] World Health Statistics 2012 (2012) Page 113.
http://apps.who.int/iris/bitstream/10665/44844/1/9789241564441_eng.pdf?ua=1

[17] International Diabetes Federation, (2014) Diabetes: Facts and Figures.
http://www.idf.org/worlddiabetesday/toolkit/gp/facts-figures

[18] World Bank, (2014) Smoking prevalence, females (% of Adults).
http://data.worldbank.org/indicator/SH.PRV.SMOK.FE

[19] World Bank (2014) Smoking Prevalence, Males (% of Adults).
http://data.worldbank.org/indicator/SH.PRV.SMOK.MA

[20] World Bank (2014) Population (Total). http://data.worldbank.org/indicator/SP.POP.TOTL

[21] Rightdiagnosis.com (2014) Statistics by Country for Stroke. http://www.rightdiagnosis.com/s/stroke/stats-country.htm.

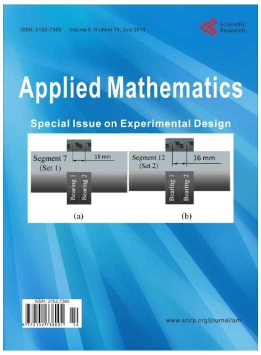
[22] R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

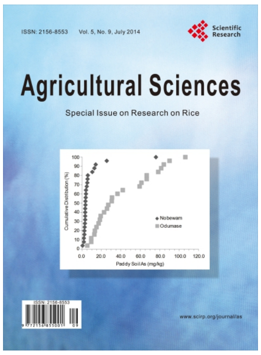
[23] Berkson, J. (1946) Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin*, **2**, 47-53. http://dx.doi.org/10.2307/3002000

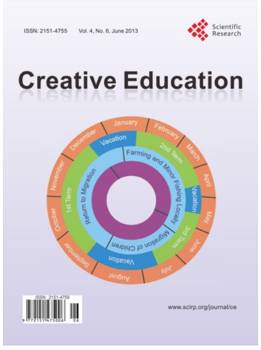
Advances in Bioscience and Biotechnology

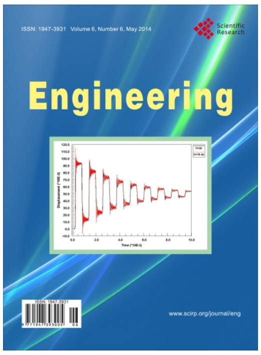
American Journal of Analytical Chemistry

American Journal of Plant Sciences
Special Issue on Photosynthesis

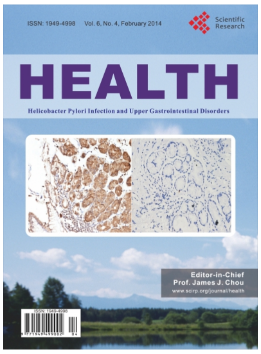

Applied Mathematics
Special Issue on Experimental Design


Agricultural Sciences
Special Issue on Research on Rice


Creative Education

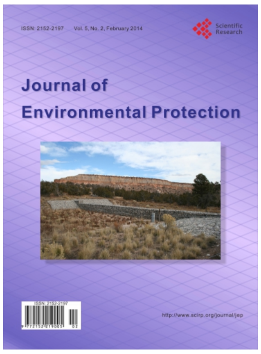

Engineering

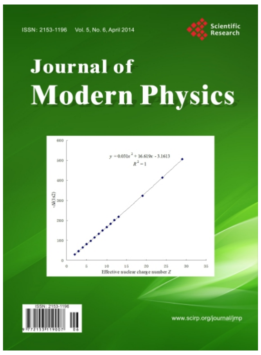
Food and Nutrition Sciences


HEALTH
Helicobacter Pylori Infection and Upper Gastrointestinal Disorders
Editor-in-Chief
Prof. James J. Chou

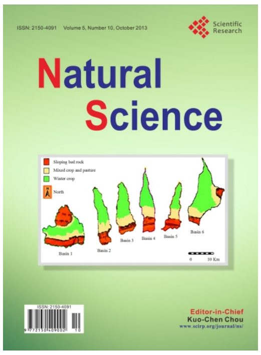
Journal of Computer and Communications

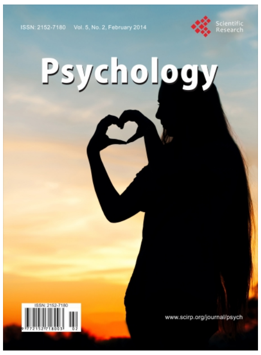
Journal of Cancer Therapy
Special Issue on Bone Tumor Research


Journal of Environmental Protection


Journal of Modern Physics

Modern Economy


Natural Science
Editor-in-Chief
Kuo-Chen Chou


Psychology