Differential loss of participants does not necessarily cause selection bias

Abstract

***Background*:** Most research is affected by differential participation, where individuals who do not participate have different characteristics to those who do. This is often assumed to induce selection bias. However, selection bias only occurs if the exposure-outcome association differs for participants compared to non-participants. We empirically demonstrate that selection bias does not necessarily occur when participation varies in a study.

***Methods*:** We used data from three waves of the longitudinal Survey of Family, Income and Employment (SoFIE). We examined baseline associations of labour market activity and education with self-rated health using logistic regression in five participation samples: A) the original sample at year one (n=22,260); B) those remaining in the sample (n=18,360); C) those (at year 3) consenting to data linkage (n=14,350); D) drop outs over three years (n=3,895); and E) those who dropped out or did not consent (n=7,905).

***Results*:** Loss to follow-up was more likely among lower socioeconomic groups and those with poorer health. However, for labour market activity and education, the odds of reporting fair/poor health were similar across all samples. Comparisons of the mutually exclusive samples (C and E) showed no difference in the odds ratios after adjustment for sociodemographic (participation) variables. Thus, there was little evidence of selection bias.

***Conclusions*:** Differential loss to follow-up (drop out) need not lead to selection bias in the association between exposure (labour market activity and education) and outcome (self-rated health).

***Key words*:** selection bias, non-response, survey data

Aust NZ J Public Health. 2012; 36:218-22
doi: 10.1111/j.1753-6405.2012.00867.x

Kristie N. Carter, Fiona Imlach-Gunasekara, Sarah K. McKenzie, Tony Blakely

Department of Public Health, University of Otago, New Zealand

All studies suffer from non-participation in some form, be it due to missing data, initial non-response or, in longitudinal studies, loss to follow-up or attrition (caused by difficulty locating participants, refusals to continue, or death), which may lead to selection bias.[1-4] Selection bias arises when the association between the exposure and outcome is different among those who participate, compared to those who do not.[5] It has been shown that non-participation (defined as non-response and attrition) more often occurs in younger populations, people of lower socioeconomic position, less stable family or household type and those in poorer health.[6,7] This will lead to biased estimates of population prevalence of sociodemographic and health characteristics.[6,8,9] However, this does not necessarily cause selection bias of the association between the exposure and outcome, as is often argued in the literature (and through peer review).[10,11] Therefore, for selection bias to occur, we would need to observe differential participation by the joint distribution of the exposure and the outcome (i.e. exposure and outcome are dependent predictors of participation).

It is often accepted that non-response and attrition in a survey automatically leads to selection bias and jeopardises the validity of results.[10,11] However, most studies only compare the characteristics of responders with non-responders and do not investigate whether any difference in participation affects the exposure and outcome association of interest.[7,12,13] In the few studies where this has been investigated, even if there is differential participation this does not lead to selection (or non-response) bias in prevalence rates[4,14,15] or baseline associations with future mortality.[6] In a comparison of attrition in the English Longitudinal Study of Ageing and the US Health and Retirement Study, it was found that although there was differential attrition between the two surveys, this had no impact on the association between different health states and socioeconomic status.[4]

The objective of this paper is to demonstrate empirically, using longitudinal data, that selection bias need not occur for the analytical association of interest, in the presence of differential participation and consent. We do so by examining the association between socioeconomic variables and self-rated health (SRH) at Wave 1 of a longitudinal study, for: A) Wave 1 original sample members; B) the balanced panel (those who participated in Waves 1, 2 and 3); C) the balanced panel restricted to those also consenting to data linkage in Wave 3; D) those lost to follow up (or dropped out) of the survey by Wave 3; and E) those who dropped out or did not consent. We do not attempt to analyse factors of non-response or time-varying attrition in this article.

Methods

Data

The Survey of Families, Income and Employment (SoFIE) is a longitudinal panel survey, administered by Statistics New Zealand (NZ), of approximately 11,500

Submitted: June 2011 **Revision requested**: September 2011 **Accepted**: November 2011
Correspondence to: Dr Kristie Carter, Department of Public Health, University of Otago, PO Box 7343, Wellington, New Zealand; e-mail: kristie.carter@otago.ac.nz

Table 1: Demographic and socioeconomic characteristics of the original (Wave 1) and restricted adult SoFIE participants.

	(A) Participants in Wave 1		(B) Balanced Panel W1,2,3			(C) Balanced Panel + Consent			(D) Drop out W2,3			(E) Drop out W2,3 + No Consent		
	N	col%	N	col%	W1%	N	col%	W1%	N	col%	W1%	N	col%	W1%
All	22,260		18,360		82.5	14,350		64.5	3,895		17.5	7,905		35.5
Sex														
Female	11,880	53.4	9,925	54.0	83.5	7,775	54.2	65.4	1,955	50.2	16.5	4,105	51.9	34.6
Male	10,380	46.6	8,440	46.0	81.3	6,575	45.8	63.3	1,940	49.8	18.7	3,800	48.1	36.6
Age														
15-24	3,890	17.5	2,755	15.0	70.8	2,130	14.8	54.8	1,135	29.1	29.2	1,760	22.3	45.2
25-34	3,735	16.8	2,950	16.1	79.0	2,300	16.0	61.6	785	20.2	21.0	1,430	18.1	38.3
35-44	4,530	20.4	3,905	21.3	86.2	3,015	21.0	66.6	625	16.0	13.8	1,520	19.2	33.6
45-54	3,760	16.9	3,340	18.2	88.8	2,575	17.9	68.5	415	10.7	11.0	1,185	15.0	31.5
55-64	2,900	13.0	2,585	14.1	89.1	2,075	14.5	71.6	315	8.1	10.9	825	10.4	28.4
65+	3,445	15.5	2,830	15.4	82.1	2,260	15.7	65.6	620	15.9	18.0	1,185	15.0	34.4
Prioritised Ethnicity														
NZ/European	16,045	72.1	13,950	76.0	86.9	11,275	78.6	70.3	2,095	53.8	13.1	4,775	60.4	29.8
Māori	3,005	13.5	2,195	12.0	73.0	1,645	11.5	54.7	810	20.8	27.0	1,360	17.2	45.3
Pacific	1,330	6.0	840	4.6	63.2	525	3.7	39.5	490	12.6	36.8	805	10.2	60.5
Asian	1,355	6.1	950	5.2	70.1	600	4.2	44.3	405	10.4	29.9	755	9.6	55.7
Other	515	2.3	425	2.3	82.5	305	2.1	59.2	90	2.3	17.5	210	2.7	40.8
Self-Rated Health														
Excellent	8,395	37.7	6,935	37.8	82.6	5,420	37.8	64.6	1,460	37.5	17.4	2,975	37.6	35.4
Very Good	7,165	32.2	6,080	33.1	84.9	4,810	33.5	67.1	1,080	27.7	15.1	2,355	29.8	32.9
Good	4,615	20.7	3,775	20.6	81.8	2,915	20.3	63.2	840	21.6	18.2	1,700	21.5	36.8
Fair/Poor	2,085	9.4	1,570	8.5	75.3	1,205	8.4	57.8	515	13.2	24.7	880	11.1	42.2
Labour Market Activity														
Unemployed	605	2.7	415	2.3	68.6	325	2.3	53.7	195	5.0	32.2	280	3.5	46.3
Inactive	8,040	36.1	6,295	34.3	78.3	4,885	34.0	60.8	1,750	44.9	21.8	3,155	39.9	39.2
Working	13,590	61.1	11,645	63.4	85.7	9,135	63.7	67.2	1,945	49.9	14.3	4,460	56.4	32.8
Education														
Degree or Higher	2,875	12.9	2,425	13.2	84.3	1,960	13.7	68.2	450	11.6	15.7	920	11.6	32.0
Post school	7,125	32.0	6,070	33.1	85.2	4,820	33.6	67.6	1,055	27.1	14.8	2,305	29.2	32.4
School Qual	6,190	27.8	5,025	27.4	81.2	3,930	27.4	63.5	1,165	29.9	18.8	2,260	28.6	36.5
No Qual	6,055	27.2	4,830	26.3	79.8	3,635	25.3	60.0	1,220	31.3	20.1	2,415	30.6	39.9
NZ Deprivation Index														
Q1 (least dep)	4,065	18.3	3,580	19.5	88.1	2,840	19.8	69.9	485	12.5	11.9	1,220	15.4	30.0
Q2	4,325	19.4	3,750	20.4	86.7	2,990	20.8	69.1	570	14.6	13.2	1,335	16.9	30.9
Q3	3,765	16.9	3,180	17.3	84.5	2,455	17.1	65.2	590	15.1	15.7	1,310	16.6	34.8
Q4	5,055	22.7	4,065	22.1	80.4	3,210	22.4	63.5	990	25.4	19.6	1,845	23.3	36.5
Q5 (most dep)	5,045	22.7	3,785	20.6	75.0	2,855	19.9	56.6	1,260	32.3	25.0	2,195	27.8	43.5

All numbers of participants presented in the tables of this paper are rounded to the nearest multiple of five, with a minimum value of 5, as per Statistics NZ confidentiality protocol, so totals may not add up to the sum of counts.

households (77% initial response rate) with more than 22,000 adults (≥15 years) interviewed on an annual basis, starting in October 2002. Annual face-to-face interviews collected comprehensive information on demographics, households, income, employment, education and family composition, as well as SRH. In Wave 3, written consent was requested from participants to link their SoFIE record to cancer registrations and hospitalisations.

Analyses

The current analyses utilise the first three waves of SoFIE data (Wave 1 to 7 data Version 1).

Cross-tabulations investigate the prevalence of demographic and socioeconomic variables in the four population restrictions (described above). To examine the effects of selection bias on the results, univariate and multivariable logistic regression analyses are used to examine the association of baseline (Wave 1) socioeconomic variables (labour market activity and education) with SRH in the four populations. SRH is dichotomised into good (excellent, very good, good) and poor (fair, poor) health. Multivariable analyses are adjusted for age, sex, ethnicity, and other socioeconomic factors (education, labour market activity and area deprivation). The Wald test is used to test for heterogeneity between the final model estimates in the mutually exclusive populations (Wave 3 responders and consenters [C] compared to Wave 3 drop out and non-consenters [E]).

All analyses are conducted on unit level data using SAS 8.2. All numbers of participants presented in the tables of this paper are rounded to the nearest multiple of five, with a minimum value of five, as per Statistics NZ confidentiality protocol, therefore totals may not add up to the sum of counts.

Results

A total of 22,260 adult original sample members participated at Wave 1 (Figure 1). By Wave 3, 18,360 (82.5%) of the original sample members were re-interviewed in Waves 2 and 3. Therefore, 3,895 participants (17.5%) dropped out of, or did not respond, in Waves 2 and/or 3. Approximately 150 deaths occurred each year, which are included in the attrition numbers. Table 1 shows that attrition was greater in younger participants, those reporting ethnicity other than NZ European, poorer health status and lower socioeconomic status (unemployed, living in highly deprived areas). Of the 18,360 people who were interviewed at Wave 3, 14,350 (78.1%) consented to having their health records linked to their SoFIE records. This represents 64.5% of the original Wave 1 population.

Figure 1: Sample flow of participants in the SoFIE survey.

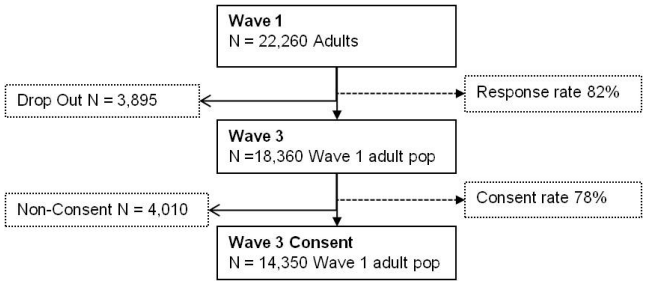

Table 2 presents the results of logistic regression analyses, regressing labour market activity and education on SRH in populations with different participation levels. For labour market activity and all levels of educational qualifications the odds ratios of reporting fair/poor health compared to good health were similar for the original Wave 1 population, the balanced panel, and the Wave 3 consenters. For example, the univariate odds of fair/poor SRH for those not working were 4.9 (95%CI 4.5-5.5) in the original Wave 1 population, 4.6 (95%CI 4.1-5.2) in the balanced panel and 4.6 (95%CI 4.0-5.2) in those who remained in the sample and consented to record linkage at Wave 3. Adjusting for age, sex, ethnicity and socioeconomic factors in the multivariable analysis reduced the associations between the socioeconomic variables and SRH.

Table 3 presents the results of the Wald test comparing the odds ratio for poor health in the most extreme (mutually exclusive) population groups: the balanced panel and consenting at Wave 3 population (C), with the population that dropped out or didn't consent (E). The odds ratios for the two population subgroups were not statistically significantly different from each other (p=0.09). Once demographic and socioeconomic factors (potentially predicting drop out) were adjusted for the odds ratios more or less identical (p=0.72).

Discussion

In this analysis of three years of longitudinal data, we have shown that people who continue to participate have different characteristics to those who drop out or do not consent to data linkage. By Wave 3 of SoFIE, 17.5% of participants had dropped out, or did not respond in Waves 2 or 3 of the survey leading to a population that is older, more likely to be of NZ European ethnicity, has better health and higher socioeconomic status (higher income, employed, living in less deprived areas). This is consistent with other research that has found those consenting to participate in research and those who continue to respond to a survey differ to those who do not.[13,16-18] However, despite this differential participation, we found little evidence of selection bias due to drop out or consent on the association between baseline socioeconomic measures and health, especially after adjustment for factors associated with participation, demographic and socioeconomic. Other studies that have looked at the effect of non-participation or attrition on regression estimates have also found minimal impact on models of exposure-outcome associations.[4,6,15,17,19,20]

The odds ratios in our study became even more similar after adjusting for covariates as these covariates were possibly predictors of participation. This is consistent with adjusting for selection bias arising due to common causes of exposure and participation, and common causes of outcome and participation (as opposed to exposure and outcome directly influencing participation), and

that adjustment for these common causes (or their proxies) will minimise any bias.[3,5]

In this analysis we do not attempt to analyse the initial household sampling non-response (23%). The SoFIE study was conducted by Statistics New Zealand, which is reflected in the high household response rate.[21] During the survey, Statistics New Zealand made great attempts to track all original sample members. If they refused follow-up or could not be found and were not interviewed for two or more consecutive years then they were no longer tracked, leading to the increasing attrition (drop-out from the sample over time). We do not attempt to examine selection bias due to time-dependent attrition or patterns of missing data in this paper.

A number of longitudinal surveys have shown that the effect of time-varying attrition on longitudinal estimates is minimal.[4,18-20,22] Some types of longitudinal analysis, such as fixed effects models, only use within individual changes over time to compute estimates so may be less prone to selection bias.

In conclusion, the use of longitudinal data allows us to examine

Table 2: Logistic regression of the relationship between Wave 1 fair/poor self-rated health and socioeconomic variables for the Wave 1 adult population, the balanced panel and those who consented to data linkage.

(A) Participants in Wave 1 N=22,260	Odds Ratio Univariate Fair/Poor	Odds Ratio Age/Sex/Eth Fair/Poor	Odds Ratio Multivariable* Fair/Poor
Labour Market Activity			
Working	1	1	1
Not Working	4.9 (4.5-5.5)	4.7 (4.2-5.3)	4.1 (3.6-4.6)
Highest education qualification			
Degree or Higher	1	1	1
Post School Qualification	2.3 (1.9-2.8)	1.9 (1.6-2.4)	1.6 (1.3-2.0)
School Qualification	1.9 (1.5-2.4)	1.9 (1.6-2.4)	1.5 (1.2-1.9)
No Qualification	4.6 (3.8-5.7)	3.2 (2.6-3.9)	2.1 (1.7-2.9)
(B) Balanced Panel W1-3 N=18,360	**Odds Ratio Univariate Fair/Poor**	**Odds Ratio Age/Sex/Eth Fair/Poor**	**Odds Ratio Multivariable* Fair/Poor**
Labour Market Activity			
Working	1	1	1
Not Working	4.6 (4.1-5.2)	4.7 (4.1-5.4)	4.1 (3.6-4.7)
Highest education qualification			
Degree or Higher	1	1	1
Post School Qualification	2.2 (1.7-2.7)	1.9 (1.5-2.4)	1.6 (1.2-2.0)
School Qualification	1.9 (1.5-2.4)	2.0 (1.5-2.5)	1.5 (1.2-2.0)
No Qualification	4.3 (3.4-5.4)	3.1 (2.5-4.0)	2.1 (1.6-2.6)
(C) Balanced Panel + Consent N=14,350	**Odds Ratio Univariate Fair/Poor**	**Odds Ratio Age/Sex/Eth Fair/Poor**	**Odds Ratio Multivariable* Fair/Poor**
Labour Market Activity			
Working	1	1	1
Not Working	4.6 (4.0-5.2)	4.7 (4.0-5.4)	4.0 (3.5-4.7)
Highest education qualification			
Degree or Higher	1	1	1
Post School Qualification	2.1 (1.6-2.8)	1.8 (1.4-2.4)	1.5 (1.2-2.0)
School Qualification	1.9 (1.4-2.4)	1.9 (1.5-2.5)	1.5 (1.1-2.0)
No Qualification	4.3 (3.3-5.5)	3.2 (2.4-4.1)	2.1 (1.6-2.8)

** adjusting for age, sex, ethnicity and socioeconomic factors (labour market activity, education and area deprivation)*

Note: these are not mutually exclusive populations

All numbers of participants presented in the tables of this paper are rounded to the nearest multiple of five, with a minimum value of 5, as per Statistics NZ confidentiality protocol, so totals may not add up to the sum of counts.

Table 3: Logistic regression of the relationship between Wave 1 fair/poor self-rated health and labour force non-participation for the balanced panel and consenting population at Wave 3 and those who dropped out or didn't consent.

	Odds Ratio Univariate Fair/Poor	Wald Test *p*-value	Odds Ratio Multivariable* Fair/Poor	Wald Test *p*-value
Balanced panel + consented; (C) (n=14,350)	4.6 (4.0-5.2)	0.0957	4.0 (3.5-4.7)	0.7154
Drop out and non-consenters; (E) (n=7,905)	5.3 (4.5-6.3)		4.2 (3.5-5.2)	
Full data Wave 1 population (n=22,260)	4.9 (4.5-5.5)		4.1 (3.0-3.9)	

All numbers of participants presented in the tables of this paper are rounded to the nearest multiple of five, with a minimum value of 5, as per Statistics NZ confidentiality protocol, so totals may not add up to the sum of counts.

the effect of non-response, attrition and consent to data linkage on the association between baseline socioeconomic factors and SRH. Although others have shown theoretically and empirically that differential participation has minimal effect on exposure-outcome associations, it is still common practice for researchers to make ill-considered assertions about selection bias based only on cross-sectional participation and differences in participation by only the exposure and the outcome separately. These results are valid for the SoFIE population only and for cross-sectional associations between labour market activity, education and health. We hope that this paper will encourage researchers to explicitly consider this bias in exposure-outcome associations, and to extend beyond presenting only considering univariate participation as an assessment of selection bias.

Acknowledgements

SoFIE-Health is primarily funded by the Health Research Council of New Zealand as part of the Health Inequalities Research Programme.Access to the data used in this study was provided by Statistics New Zealand in a secure environment designed to give effect to the confidentiality provisions of the Statistics Act, 1975. The results in this study and any errors contained therein are those of the author, not Statistics New Zealand.

References

1. Delgado-Rodríguez M, Llorca J. Bias. *J Epidemiol Community Health*. 2004;58:635-41.
2. Kleinbaum DG, Morgenstern H, Kupper L. Selection bias in epidemiologic studies. *Am J Epidemiol*. 1981;113(4):452-63.
3. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615-25.
4. Banks J, Murieal A, Smith JP. Attrition and health in ageing studies: evidence from ELSA and HRS. *Longit Life Course Stud*. 2011;2(2):101-26.
5. Rothman KJ, Greenland S, Lash TL, editors. *Modern Epidemiology*. 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008.
6. Batty GD, Gale CR. Impact of resurvey non-response on the associations between baseline risk factors and cardiovascular disease mortality: prospective cohort study. *J Epidemiol Community Health*. 2009;63:952-5.
7. Lorant V, Demarest S, Miermans P-J, Van Oyen H. Survey error in measuring socio-economic risk factors of health status: a comparison of a survey and a census. *Int J Epidemiol*. 2007;36(6):1292-9.
8. Buckley B, Murphy AW, Byrne M, Glynn L. Selection bias resulting from the requirement for prior consent in observational research: a community cohort of people with ischaemic heart disease. *Heart*. 2007;93(9):1116-20.
9. Ives DG, Traven ND, Kuller LH, Schulz R. Selection Bias and Nonresponse to Health Promotion in Older Adults. *Epidemiology*. 1994;5(4):456-61.
10. Vestbo J, Rasmussen FV. Baseline characteristics are not sufficient indicators of non-response bias follow up studies. *J Epidemiol Community Health*. 1992;46(6):617-9.
11. Sheikh K. Investigation of selection bias using inverse probability weighting. *Eur J Epidemiol*. 2007;22(5):349-50.
12. Strandhagen E, Berg C, Lissner L, Nunez L, Rosengren A, Torén K, et al. Selection bias in a population survey with registry linkage: potential effect on socioeconomic gradient in cardiovascular risk. *Eur J Epidemiol*. 2010;25(3):163-72.
13. Contoyannis P, Jones AM, Rice N. The dynamics of health in the British Household Panel Survey. *J Appl Econom*. 2004;19(4):473-503.
14. Mannetje At, Eng A, Douwes J, Ellison-Loschmann L, McLean D, Pearce N. Determinants of non-response in an occupational exposure and health survey in New Zealand. *Aust N Z J Public Health*. 2011;35(3):256-63.
15. de Winter A, Oldehinkel A, Veenstra R, Brunnekreef J, Verhulst F, Ormel J. Evaluation of non-response bias in mental health determinants and outcomes in a large sample of pre-adolescents. *Eur J Epidemiol*. 2005;20(2):173-81.
16. Søgaard AJ, Selmer R, Bjertness E, Thelle D. The Oslo Health Study: The impact of self-selection in a large, population-based survey. *Int J Equity Health*. 2004;3:3-12.
17. Alonso A, Seguí-Gómez M, de Irala J, Sánchez-Villegas A, Beunza J, Martínez-Gonzalez M. Predictors of follow-up and assessment of selection bias from dropouts using inverse probability weighting in a cohort of university graduates. *Eur J Epidemiol*. 2006;21(5):351-8.
18. Jones AM, Koolman X, Rice N. Health-related non-response in the British Household Panel Survey and European Community Household Panel: using inverse-probability-weighted estimators in non-linear models. *J R Stat Soc Ser A Stat Soc*. 2006;169(3):543-69.
19. Powers J, Loxton D. The Impact of Attrition in an 11-Year Prospective Longitudinal Study of Younger Women. *Ann Epidemiol*. 2010;20(4):318-21.
20. Howe LD, Galobardes B, Tilling K, Lawlor DA. Does drop-out from cohort studies bias estimates of socioeconomic inequalities in health? IEA World Congress of Epidemiology; 2011; Edinburgh. *J Epidemiol Community Health.* 2011;65:A31.
21. Carter KN, Cronin M, Blakely T, Hayward M, Richardson K. Cohort Profile: Survey of Families, Income and Employment (SoFIE) and Health Extension (SoFIE-health). *Int J Epidemiol*. 2010;39(3):653-9.
22. Lillard LA, Panis CWA. Panel attrition from the panel study of income dynamics − Household income, marital status, and mortality. *J Hum Resour*. 1998;33(2):437-57.