On the Impact of Informative Nonresponse in Logistic Regression

Joanna J.J. Wang* Mark Bartlett† Louise Ryan‡

Abstract

In this paper, we are interested in nonignorable missing data mechanism where the probability of nonresponse depends on the outcome. We consider a selection model for nonignorable nonresponse in logistic regression. Expressions for the bias in parameter estimates are derived in a simple case. Further, we propose a sensitivity analysis to study changes in parameter estimates under different assumptions. We adopt a Bayesian framework as it offers a flexible approach for incorporating different missing data mechanisms. Our modelling strategy is illustrated using survey data from the 45 and Up Study.

Key Words: 45 and Up study, informative nonresponse; selection model, sensitivity analysis

1. Introduction

Missing data and nonresponse are common in epidemiological studies and they pose major methodological challenges. Missing data can result in a loss of statistical power and an increase in variances of estimates due to the loss of observations. Nonresponse can also induce bias in estimates since responders to follow-up surveys may have very different attributes to nonresponders (Nohr et al. 2006; Young et al. 2006).

One popular approach to compensate for attrition in longitudinal studies is inverse propensity score weighting (Rosenbaum, 1987) where weights derived from response probabilities are assigned to the responders to ensure the distribution of the original population is properly represented. A second commonly used approach for handling missing data is multiple imputation, which generates multiple sets of imputed values for missing observations from suitable probability distributions. Both methods generally rely on the "missing at random" (MAR) assumption, which asserts that missingness depends only on the observed information. However, in many situations, the process that generates missingness may be directly related to the values of the unobserved variables. For example, in the longitudinal study that motivates this paper, it is reasonable to think that baseline survey participants who moved to a new dwelling-type during the follow-up period may be less likely to respond to the follow-up survey. In such cases, assuming that the data are MAR may yield biased results. A number of authors in recent years have proposed strategies to handle this informative missingness. One of the most popular models is a selection model (Diggle and Kenward, 1994) that combines a linear model for the outcome and a logistic regression model for the missingness process. This class of models have also been explored by Scharfstein et al. (1999), Ibrahim et al. (2001) and Carpenter et al. (2002), among many others. In this paper, we propose a selection model for nonresponse in logistic regression. We derive expressions for the bias in regression parameters and propose a Bayesian sensitivity analysis for nonignorable missingess.

We illustrate our modelling strategy with survey data from the Sax Institute's 45 and Up Study. This is the largest cohort study of population aging even undertaken in Aus-

*School of Mathematical and Physical Sciences, University of Technology Sydney, Australia. The Sax Institute, Sydney, Australia. The Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)

†The Sax Institute, Sydney, Australia. ACEMS

‡School of Mathematical and Physical Sciences, University of Technology Sydney, Australia. ACEMS

tralia. Recruitment into the 45 and Up Study commenced in 2006 and the first follow-up survey began in 2012 and continued for 4 years. The cohort consists of more than 267,000 men and women aged 45 years and over from the general population of the state of New South Wales (NSW), Australia. Extensive information was collected on demographic and social-economic characteristics; personal health behaviours and general health related data including known risk factors for major causes of morbidity and mortality.

2. Models

In this section, we first derive expressions of the bias in regression parameters in a simple case of logistic regression. We then present a Bayesian selection model and propose a straightforward method for preforming a sensitivity analysis.

2.1 A simple logistic regression model

Consider a simple case of logistic regression with binary outcome y_i and a single covariate x_i for subject i. Let $\pi_i = P(y_i = 1|x_i)$ and $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$. This implies

$$\pi_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))}.$$

Let m_i be the missingness indicator variable taking the value of 1 if y_i is missing and 0 otherwise. Let $p_i = P(m_i = 1|y_i, x_i)$ and $\text{logit}(p_i) = \alpha_0 + \alpha_1 x_i + \lambda y_i$.

If we perform a complete case analysis, then inference is based on observed data likelihood defined as

$$\begin{aligned} L(\beta_0, \beta_1|y_i, x_i) &= \prod_{i=1}^{n} [\pi^{y_i}(1-\pi_i)^{1-y_i}]^{(1-m_i)} \\ &= \prod_{i=1}^{n} \left[\left(\frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))} \right)^{y_i} \left(1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))} \right)^{1-y_i} \right]^{(1-m_i)} \\ &= \prod_{i=1}^{n} \left[\frac{\exp(y_i(\beta_0 + \beta_1 x_i))}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{(1-m_i)}. \end{aligned}$$

Thus the log-likelihood is

$$l(\beta_0, \beta_1|y_i, x_i) = \sum_{i=1}^{n} y_i(1-m_i)(\beta_0 + \beta_1 x_i) - (1-m_i)\ln(1 + \exp(\beta_0 + \beta_1 x_i)).$$

Partially differentiate with respect to β_1, we get

$$\begin{aligned} \frac{\partial l(\beta_0, \beta_1|y_i, x_i)}{\partial \beta_1} &= \sum_{i=1}^{n} x_i y_i (1-m_i) - \frac{x_i(1-m_i)}{1 + \exp(-(\beta_0 + \beta_1 x_i))} \\ &= \sum_{i=1}^{n} x_i y_i (1-m_i) - (1-m_i)\pi_i x_i. \end{aligned}$$

Taking expectation, we have

$$\begin{aligned} E\left(\frac{\partial l(\beta_0, \beta_1|y_i, x_i)}{\partial \beta_1} \right) &= \sum_{i=1}^{n} x_i P(y_i = 1|m_i = 0, x_i) P(m_i = 0|x_i) - \pi_i x_i (1-p_i) \\ &= \sum_{i=1}^{n} x_i P(y_i = 1|m_i = 0, x_i)(1-p_i) - \pi_i x_i (1-p_i) \end{aligned}$$

$$= \sum_{i=1}^{n} \pi_i^*(1-p_i) - \pi_i x_i(1-p_i),$$

where

$$\pi_i^* = P(y_i = 1 | m_i = 0, x_i)$$

which is the probability of $y_i = 1$ in the complete case analysis.

Let

$$\begin{aligned} p_i(1) &= P(m_i = 1 | y_i = 1, x_i), \\ p_i(0) &= P(m_i = 1 | y_i = 0, x_i). \end{aligned}$$

To derive the bias of regression coefficients using observed cases only, consider

$$\begin{aligned} \text{logit}(\pi_i^*) &= \ln\left(\frac{\pi_i(1-p_i(1))}{1-p_i} \times \frac{1-p_i}{(1-\pi_i)(1-p_i(0))}\right) \\ &= \ln \pi_i - \ln(1-\pi_i) + \ln\left(\frac{1-p_i(1)}{1-p_i(0)}\right) \\ &= \beta_0 + \beta_1 x_i + \ln\left(\frac{1-p_i(1)}{1-p_i(0)}\right) \\ &= \beta_0 + \beta_1 x_i + \ln\left(\frac{1+\exp(\alpha_0+\alpha_1 x_i)}{1+\exp(\alpha_0+\alpha_1 x_i+\lambda)}\right). \end{aligned} \tag{1}$$

If the covariate x_i is binary, so (1) becomes

$$\begin{aligned} \text{logit}(\pi_i^*) &= (\beta_0 + \Delta_0) + (\beta_1 + (\Delta_1 - \Delta_0))x_i \\ &= \beta_0^* + \beta_1^* x_i, \end{aligned}$$

where

$$\begin{aligned} \Delta_0 &= \ln\left(\frac{1+\exp(\alpha_0)}{1+\exp(\alpha_0+\lambda)}\right) \qquad \text{if } x_i = 0 \\ \Delta_1 &= \ln\left(\frac{1+\exp(\alpha_0+\alpha_1)}{1+\exp(\alpha_0+\alpha_1+\lambda)}\right) \quad \text{if } x_i = 1, \end{aligned} \tag{2}$$

and β_0^* and β_1^* are parameters using complete cases only.

Hence the parameters using the full data are

$$\begin{aligned} \beta_0 &= \beta_0^* - \Delta_0, \\ \beta_1 &= \beta_1^* - (\Delta_1 - \Delta_0). \end{aligned} \tag{3}$$

The terms Δ_0 and $(\Delta_1 - \Delta_0)$ represent the bias of the intercept and regression coefficient respectively, resulting from using complete cases only. This result shows that if the covariate x_i is a weak predictor for response missingness so that the coefficient α_1 is close to 0, then the bias $(\Delta_1 - \Delta_0)$ in the regression coefficient goes to 0.

2.2 A Bayesian selection model

We present a Bayesian selection model to account for nonresponse which we assume is "not missing at random" (NMAR). In this case, the missing data mechanism must be specified by the researcher and incorporated into the model to obtain unbiased parameter estimates.

However, observed data contain no information about the missing data mechanism and statistical inference is very sensitive to the choice of such formulation. This makes sensitivity analysis essential for investigating possible violations of the MAR assumption and exploring the robustness of the study conclusions to increasingly extreme departures from the MAR mechanism (Verbeke et al. 2001; Scharfstein et al. 2003).

A selection model consists of two sub-models: one specifies the relationship between the covariates and the outcome of interest and the other represents the missing data process, which is dependent not only on observed covariates, but also the outcome. More specifically, we assume a standard logistic regression for the binary outcome of interest:

$$\text{logit}(P(y_i = 1)) = \beta_0 + \sum_{j=1}^{k} b_j x_{ji}, \tag{4}$$

where y_i is the outcome and x_{ji} is the jth covariate for subject i. We then add a logistic model for nonresponse as follows:

$$\text{logit}(P(m_i = 1)) = \theta_0 + \sum_{s=1}^{l} \theta_s x_{si} + \lambda y_i, \tag{5}$$

where m_i is a nonresponse indicator defined before.

Equation (5) specifies a linear relationship between the logit of the probability of nonresponse and the outcome. Different values of the parameter λ posit different assumptions on how strongly the likelihood of nonresponse depends on the outcome. A special case is when $\lambda = 0$ which corresponds to the MAR assumption. This parameter is interpreted as the log odds ratio (OR) of nonresponse for those who had the outcome of interest. In implementing the selection model, we repeat the analysis for a range of values of λ and examine the sensitivity of the estimated regression coefficients in the outcome equation (4).

A full Bayesian probability modelling approach using Markov chain Monte Carlo (MCMC) was used for the selection model, as it was shown that the Bayesian modelling approach provides a flexible way to incorporate different assumptions on the missing data mechanism and enables coherent model estimation (Mason et al. 2010). We ran the Bayesian selection model in WinBUGS (Lunn et al. 2000) for 15,000 iterations including 5000 burn-in. Vague $N(0, 1000)$ prior distributions were assigned to intercept parameters β_0 and θ_0 and all regression coefficients in equations (4) and (5). Visual inspection of trace plots and autocorrelation plots of MCMC iterations revealed satisfactory convergence.

3. Application

The Sax Institute's 45 and Up Study is a population-based sample from the state of NSW where prospective participants were randomly sampled from the Department of Human Services enrolment database. Recruitment commenced in February 2006 and the full cohort of size 267,157 reached by December 2009. Detailed description of the 45 and Up Study can be found in 45 and Up Study Collaborators (2008).

The first wave of follow-up of participants began in 2012 with 41,440 Study participants invited. Of these, 27,036 returned the follow-up questionnaire, resulting in a response rate of 65.2%. After excluding individuals with missing values for baselines covariates, 32,037 individuals were included in this analysis with 21,750 of these being responders to the follow-up survey.

The conduct of the 45 and Up Study was approved by the University of New South Wales Human Research Ethics Committee (HREC).

The binary outcome of interest we choose is whether there was a dwelling-type change between the baseline and follow-up survey. Potential covariates for the outcome, as well as those that may be predictors for nonresponse are detailed in Table A in the Appendix.

The 45 and Up Study baseline and follow-up questionnaire ask respondents to describe their dwelling types as belonging to one of eight categories: house, flat/unit/apartment, house on farm, retirement village/self-care unit, nursing home, hostel for the aged, mobile home and other. Due to low counts in some categories of these variables, nursing home and hostel for the aged are combined into one category. Similarly, remote and very remote Accessibility/Remoteness Index of Australia (ARIA) categories are combined.

Most of the baseline variables were taken directly as responses to the relevant questions. Physical functional limitation was using the RAND 36-Item Health Survey, Version 1.0, subscale. The subscale was scored as recommended in "Scoring Instructions for MOS 36-Item Short Form Survey Instrument (SF-36)" (RAND Health, 2009). Social connectedness was assessed using the Duke Social Support Index (DSSI) subscale and scored as recommended by Broadhead et al. (1988). As per Phongsaven et al. (2013), due to the positively skewed distribution of the scores, this variable was transformed into quartiles. The change in dwelling-type was assessed by comparing responses to the relevant questions between the baseline and the follow-up survey.

In implementing the Bayesian selection model, we assume that λ is nonnegative so that the likelihood of response is higher for those who had a dwelling-type change. This is a plausible assumption since those who had dwelling-type change are more difficult to track in a longitudinal study (Voorpostel and Lipps, 2011). Furthermore, the values we choose for λ are $(0, 0.5, 1, 1.5)$. These values imply that the OR of nonresponse for individuals with dwelling-type change is between 1 and 4.5 (Uhrig, 2008; Voorpostel and Lipps, 2011).

3.1 Results

Table 1 presents the distribution of demographic and other characteristics at baseline including dwelling type, work status and carer status, self-reported health conditions, physical functional limitation and social connectedness among responders and nonresponders. After removing those with missing values in any baseline covariates listed in Appendix Table A, 67.2% of individuals responded to the follow-up survey. OR and 95% credible interval (CI) estimated from the multivariable logistic regression model for response are also presented in Table 1.

The results of modelling response showed that individuals with the following characteristics have higher odds of responding to the follow-up survey as compared with each reference category: female, in 55-74 age categories, having higher educational qualifications and having higher household income. Conversely, those who were single, worked full-time, had poor self-rated health, had moderate to severe functional limitation, had poor social connectedness, were a carer and born outside Australia are more likely to be nonresponders.

Table 2 shows the OR estimates and 95% CIs for the complete case analysis for the association between dwelling-type change and various baseline characteristics. The results showed the likelihood of having a dwelling-type change is significantly greater for individuals who were over 75 years of age, were separated at baseline, did not live in a house and not lived in major cities at baseline. On the other hand, those with household income more than $\$70,000$ were significantly less likely to have dwelling-type change between surveys.

Table 2 also includes parameter estimates from fitting Bayesian selection model for nonignorable missing mechanism. A special case of NMAR is $\lambda = 0$, which corresponds to the MAR assumption. Analysis under the MAR assumption produced very similar results

to analysing the completely observed data. Assuming the MAR assumption is plausible, it is reasonable to conclude that nonresponse is unlikely to alter the conclusions.

The sensitivity of parameter estimates to the possibility of NMAR is also examined by fitting the selection model with some plausible values of λ. As the degree of departure from MAR increases, the OR estimate for the gender variable gradually shifts away from the null value, but the 95% CIs include the null value in all cases. Including an explicit model for NMAR also had minimal impact on baseline marital status, dwelling-type and ARIA variables since the OR estimates and 95% CIs are consistent across different modelling assumptions and there is no change in the interpretation of results. More specifically, the association of dwelling-type change with baseline dwelling-type and ARIA remains significantly positive under both MAR and NMAR assumptions with different degree of nonignorability. In the complete case analysis and the selection model assuming MAR, the OR estimates for income categories $\$20,000 - \$40,000$ and $\$40,000 - \$70,000$ were of borderline significance. After allowing for NMAR, the point estimates decreased away from the null value and the CIs no longer include 1.

Forest plots showing how OR estimates and CIs change to different modelling assumptions (i.e. complete case, MAR and NMAR) and change in for significant predictors of dwelling-type change are shown in Figures 1 to 6. These plots clearly demonstrate the overall robustness of our conclusions to the possibility of MAR and NMAR assumptions.

4. Discussion

A major threat to the validity of longitudinal studies is nonresponse, which could affect the magnitude and direction of measures of association. Using the baseline and follow-up questionnaire data from the 45 and Up Study, we were able to identify a range of factors associated with response to the follow-up survey in this large cohort. More than 65% of the invited participants from the baseline responded to the follow-up survey.

Characteristics associated with a higher probability of responding to the follow-up questionnaire included: female gender, age categories 55-74, higher educational qualification, married, worked part time or partially or fully retired and higher household income. Those who were born outside Australia, who spoke a language other than English at home, were a carer, who reported poorer subjective health, who had significant functional limitation and poor social connectedness had lower odds of responding to the follow-up survey. There is no statistically significant difference in response by ARIA and most strata of baseline dwelling-type. Generally speaking, our findings on the characteristics associated with response are in accordance with many previous studies (Etter and Perneger, 1997; Watson and Wooden, 2009).

The use of a Bayesian selection model allows us to further assess the robustness of parameter estimates and conclusions when we have reasons to believe the missing data mechanism is NMAR. In implementing the Bayesian selection model, we repeated our analysis over a range of fixed values of the sensitivity parameter λ, which controls the degree of departure from the MAR assumption, as a form of sensitivity analysis. The results from the selection model indicate that for the range of λ values we considered, nonignorable nonresponse did not substantially affect estimates and conclusions for variables that were significantly associated with a dwelling-type change.

Our results also indicated that some variables are more sensitive to the underlying missing data mechanism and increasing departure from the MAR assumption. It was shown in (3) that if the covariate is a weak predictor for nonresponse, then the bias of the corresponding regression coefficient in the outcome equation diminishes. This agrees with what we observe in the real data application. For instance, baseline marital status, dwelling-type and

area remoteness were not significant predictors for response missingness and their corresponding ORs were quite robust to different assumptions on the missing data mechanism. In contrast, variables including gender, age categories and household income were significant predictor for nonresponse and their OR estimates varied substantially with different missing data assumptions and increasing values of λ.

There are several limitations in this study. First, we have assumed a linear pattern of missingness in the selection model. It may be worthwhile to explore alternate specification of the functional form. Second, it is possible that there are some unmeasured confounding factors associated with the outcome and/or nonresponse that were not captured. However, since a large number of covariates were collected at baseline, the likelihood of uncaptured confounders is low. Third, in our application we did not distinguish between different types of nonresponse. For example, reasons for nonresponse could be due to refusal or inability to be contacted. This can be accounted for by extending the model of nonresponse by using multiple missingness indicators. Also, we restrict our analysis to individuals with fully observed covariates at baseline, those with missing values in any baseline covariates could be incorporated by using methods such as multiple imputation. Lastly, we conducted the sensitivity analysis for a range of λ values which we assume to be plausible for quantifying the probability of nonresponse for individuals with and without a dwelling type change. Ideally we would want strong scientific evidence to support the use of particular values of λ.

Acknowledgements

This research was completed using data collected through the 45 and Up Study. The 45 and Up Study is managed by the Sax Institute in collaboration with major partner Cancer Council NSW; and partners: the National Heart Foundation of Australia (NSW Division); NSW Ministry of Health; NSW Government Family & Community Services - Carers, Ageing and Disability Inclusion; and the Australian Red Cross Blood Service. We thank the many thousands of people participating in the 45 and Up Study.

Table 1: Characteristics of 45 and Up Study participants according to response to follow-up survey

	Responded ($n = 21750$)		Not responded ($n = 10287$)	Total ($n = 32037$)	
Baseline characteristics	Moved ($n = 3005$)	Not moved ($n = 18745$)			OR (95% CI)
Gender					
Male	1302 (43.3)	8216 (43.8)	4736 (46.0)	14254	Ref
Female	1703 (56.7)	10529 (56.2)	5551 (54.0)	17783	1.13 (1.07-1.19)
Age (yrs)					
45-54	887 (29.5)	5917 (31.6)	3391 (33.0)	10195	Ref
55-64	1200 (39.9)	7782 (41.5)	3672 (35.7)	12654	1.21 (1.13-1.28)
65-74	590 (19.6)	3550 (18.9)	1924 (18.7)	6064	1.11 (1.01-1.21)
75-84	279 (9.3)	1328 (7.1)	1061 (10.3)	2668	0.84 (0.75-0.95)
85+	49 (1.7)	168 (0.9)	239 (2.3)	456	0.59 (0.48-0.73)
Highest qualification					
None	226 (7.6)	1334 (7.1)	1346 (13.1)	2906	Ref
Year 10	576 (19.2)	3664 (19.6)	2422 (23.5)	6662	1.22 (1.11-1.34)
Year 12	305 (10.1)	1818 (9.7)	1172 (11.4)	3295	1.41 (1.27-1.57)
Trade	267 (8.9)	1695 (9.0)	1122 (10.9)	3084	1.34 (1.20-1.49)

Cert./diploma	777 (25.9)	4490 (24.0)	2117 (20.6)	7384	1.73 (1.57-1.90)
Tertiary	854 (28.4)	5744 (30.7)	2108 (20.5)	8706	2.19 (1.98-2.41)
Area remoteness					
Major cities	1307 (43.5)	10145 (54.1)	5713 (55.5)	17165	Ref
Inner regional	1295 (43.1)	6913 (36.9)	3590 (34.9)	11798	1.05 (0.99-1.11)
Outer regional	363 (12.1)	1594 (8.5)	906 (8.8)	2863	1.05 (0.96-1.15)
Remote/very remote	40 (1.3)	93 (0.5)	73 (0.7)	211	0.87 (0.65-1.18)
Country of birth					
Australia	2332 (77.6)	14714 (78.5)	7413 (72.1)	24459	Ref
NW Europe	404 (13.4)	2368 (12.6)	1250 (12.2)	4022	0.95 (0.88-1.02)
S & E Europe	45 (1.5)	329 (1.8)	380 (3.7)	754	0.66 (0.56-0.78)
Middle East	24 (0.8)	91 (0.5)	145 (1.4)	260	0.52 (0.39-0.67)
SE Asia	20 (0.7)	191 (1.0)	252 (2.5)	463	0.46 (0.38-0.57)
NE Asia	18 (0.6)	160 (0.9)	192 (1.9)	370	0.56 (0.44-0.70)
S & Central Asia	14 (0.5)	92 (0.5)	101 (1.0)	207	0.50 (0.37-0.66)
America	41 (1.4)	189 (1.0)	146 (1.4)	376	0.63 (0.50-0.78)
Sub Saharan Africa	26 (0.9)	162 (0.9)	124 (1.2)	312	0.59 (0.46-0.74)
Oceania	79 (2.6)	436 (2.3)	279 (2.7)	794	0.77 (0.66-0.90)
Speak a language other than English at home					
No	2830 (94.2)	17575 (93.8)	8983 (87.3)	29388	Ref
Yes	175 (5.8)	1170 (6.2)	1304 (12.7)	2649	0.69(0.62-0.76)
Marital status					
Single	181 (6.0)	882 (4.7)	612 (6.0)	1675	Ref
Married	1951 (64.9)	13864 (74.0)	7069 (68.7)	22884	1.15 (1.03-1.28)
De facto	212 (7.1)	1083 (5.8)	579 (5.6)	1874	1.09 (0.94-1.26)
Widowed	187 (6.3)	1036 (5.5)	774 (7.5)	1997	1.05 (0.90-1.21)
Divorced	321 (10.7)	1427 (7.6)	918 (8.9)	2666	1.02 (0.89-1.16)
Separated	153 (5.1)	453 (2.4)	335 (3.3)	941	1.02 (0.86-1.21)
Work status					
FT/self-employed	1238 (41.2)	7895 (42.1)	4134 (40.2)	13267	Ref
PT	381 (12.7)	2849 (15.2)	1278 (12.4)	4508	1.19 (1.09-1.29)
Fully retired	953 (31.7)	5708 (30.5)	3267 (31.8)	9928	1.38 (1.26-1.50)
Partially retired	131 (4.4)	762 (4.1)	301 (2.9)	1194	1.38 (1.20-1.59)
Disabled/sick	91 (3.0)	369 (2.0)	427 (4.2)	887	1.09 (0.93-1.28)
Look after home	174 (5.8)	982 (5.2)	681 (6.6)	1837	1.07 (0.95-1.19)
Unemployed	37 (1.2)	180 (1.0)	199 (1.9)	416	0.88 (0.71-1.08)
Income category					
$< \$20,000$	607 (20.2)	2620 (14.0)	2173 (21.1)	5400	Ref
$\$20,000 - \$40,000$	572 (19.0)	3324 (17.7)	1868 (18.2)	5764	1.11 (1.02-1.21)
$\$40,000 - \$70,000$	672 (22.4)	4077 (21.8)	1931 (18.8)	6680	1.24 (1.13-1.35)
$> \$70,000$	794 (26.4)	6267 (33.4)	2511 (24.4)	9572	1.26 (1.15-1.39)
Prefer not to answer	360 (12.0)	2457 (13.1)	1804 (17.5)	4621	0.85 (0.78-0.93)
Dwelling type					
House	1558 (51.8)	15681 (83.7)	7953 (77.3)	25192	Ref
Flat/unit/apart.	538 (17.9)	1556 (8.3)	1201 (11.7)	3295	0.93 (0.85-1.01)
House on farm	654 (21.8)	1134 (6.1)	714 (6.9)	2502	1.12 (1.02-1.23)
Retirement village	58 (1.9)	249 (1.3)	180 (1.8)	487	1.04 (0.86-1.27)
Nursing home/hostel	18 (0.6)	17 (0.1)	38 (0.4)	73	0.45 (0.14-1.44)
Mobile home	72 (2.4)	68 (0.4)	81 (0.8)	221	1.02 (0.77-1.36)

Other	107 (3.6)	40 (0.2)	120 (1.2)	267	0.74 (0.57-0.95)
Carer status					
No	2639 (87.8)	16615 (88.6)	8939 (86.9)	28193	Ref
Yes	366 (12.2)	2130 (11.4)	1348 (13.1)	3844	0.90 (0.84-0.97)
Self-rated health					
Excellent	551 (18.3)	3480 (18.6)	1335 (13.0)	5366	Ref
Very good	1177 (39.1)	7629 (40.7)	3504 (34.1)	12310	0.88 (0.82-0.95)
Good	909 (30.3)	5881 (31.4)	3653 (35.5)	10443	0.74 (0.68-0.80)
Fair	316 (10.5)	1541 (8.2)	1490 (14.5)	3347	0.59 (0.53-0.65)
Poor	52 (1.8)	214 (1.1)	305 (3.0)	571	0.53 (0.43-0.65)
Functional limitation (fl)					
No fl	2149 (71.5)	14176 (75.6)	6867 (66.8)	23192	Ref
Slight fl	444 (14.8)	2559 (13.7)	1498 (14.6)	4501	1.04 (0.97-1.12)
Moderate fl	189 (6.3)	1040 (5.6)	813 (7.9)	2042	0.94 (0.85-1.05)
Significant fl	143 (4.8)	597 (3.2)	621 (6.0)	1361	0.78 (0.69-0.89)
Severe fl	80 (2.7)	373 (2.0)	488 (4.7)	941	0.71 (0.61-0.84)
Duke Social Support Index (DSSI) in quartiles					
1	1215 (40.4)	7635 (40.7)	3748 (36.4)	12598	Ref
2	697 (23.2)	4605 (24.6)	2411 (23.4)	7713	0.98 (0.92-1.05)
3	524 (17.5)	3227 (17.2)	1876 (18.2)	5627	0.95 (0.89-1.02)
4	569 (18.9)	3278 (17.5)	2252 (21.9)	6099	0.90 (0.84-0.97)

Table 2: Multivariable logistic regression analysis of characteristics associated with changed dwelling-type between baseline and follow-up surveys. Bold font indicates statistically significant results at 5% level.

Baseline characteristics	Complete case	Selection model with $\lambda = 0$
Gender		
Male	Ref	Ref
Female	0.991 (0.91-1.08)	0.989 (0.91-1.08)
Age		
45-54	Ref	Ref
55-64	0.978 (0.89-1.08)	0.976 (0.88-1.08)
65-74	1.038 (0.91-1.18)	1.034 (0.91-1.18)
75-84	**1.361 (1.15-1.62)**	**1.358 (1.14-1.62)**
85+	**1.714 (1.19-2.46)**	**1.700 (1.18-2.46)**
Marital status		
Single	Ref	Ref
Married	0.895 (0.75-1.07)	0.893 (0.74-1.07)
De facto	1.138 (0.90-1.44)	1.133 (0.90-1.43)
Widowed	0.832 (0.65-1.06)	0.828 (0.65-1.06)
Divorced	1.116 (0.90-1.38)	1.113 (0.90-1.38)
Separated	**1.763 (1.36-2.29)**	**1.757 (1.36-2.28)**
Income category		
$< \$20,000$	Ref	Ref
$\$20,000 - \$40,000$	0.882 (0.77-1.01)	0.881 (0.77-1.01)
$\$40,000 - \$70,000$	0.886 (0.77-1.02)	0.884 (0.77-1.01)
$> \$70,000$	**0.788 (0.69-0.90)**	**0.785 (0.68-0.90)**
Prefer not to answer	**0.785 (0.67-0.92)**	**0.784 (0.67-0.92)**
Dwelling type		
House	Ref	Ref
Flat/unit/apart.	**3.251 (2.89-3.66)**	**3.248 (2.89-3.66)**
House on farm	**5.388 (4.79-6.06)**	**5.398 (4.81-6.06)**
Retirement village	**1.924 (1.42-2.60)**	**1.912 (1.41-2.59)**
Nursing home/hostel	**8.755 (4.43-17.31)**	**8.837 (4.42-17.68)**
Mobile home	**8.990 (6.39-12.65)**	**9.025 (6.41-12.71)**
Other	**23.108 (15.93-33.51)**	**23.524 (16.17-34.22)**
Area remoteness		
Major cities	Ref	Ref
Inner regional	**1.227 (1.12-1.35)**	**1.227 (1.12-1.35)**
Outer regional	**1.251 (1.08-1.45)**	**1.249 (1.08-1.44)**
Remote/very remote	**1.813 (1.20-2.75)**	**1.799 (1.18-2.74)**

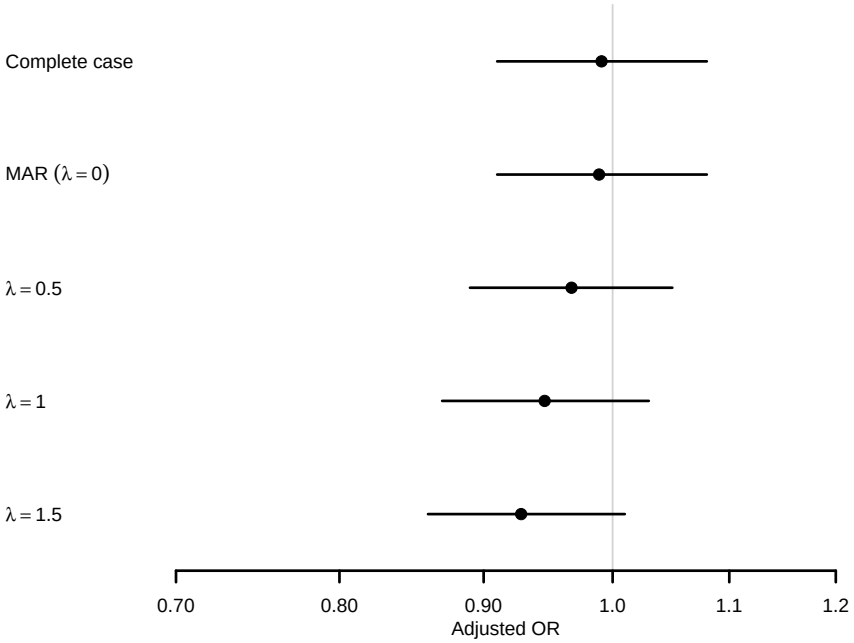

Figure 1: Estimated OR and 95% CI for gender (Ref='Male'), under the complete case and selection model with different values of λ.

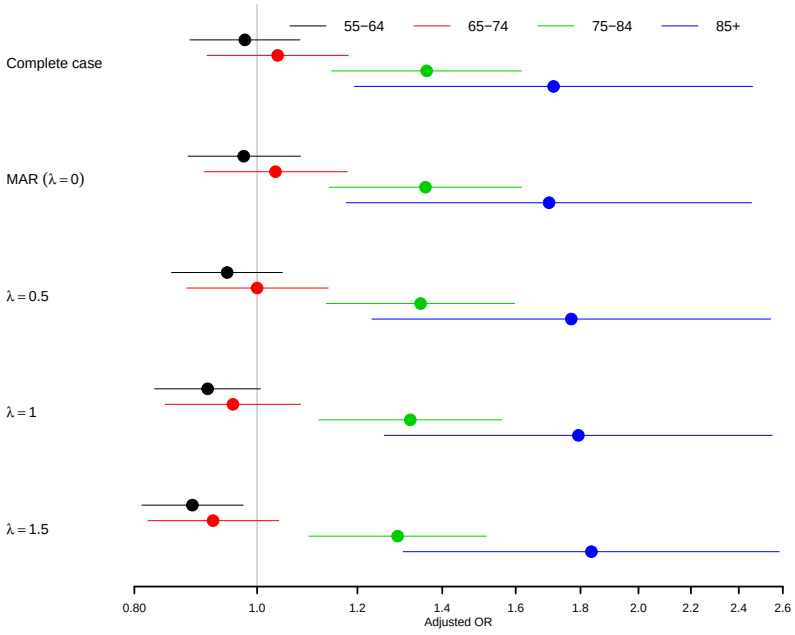

Figure 2: Estimated OR and 95% CI for age group (Ref='45-54'), under the complete case and selection model with different values of λ.

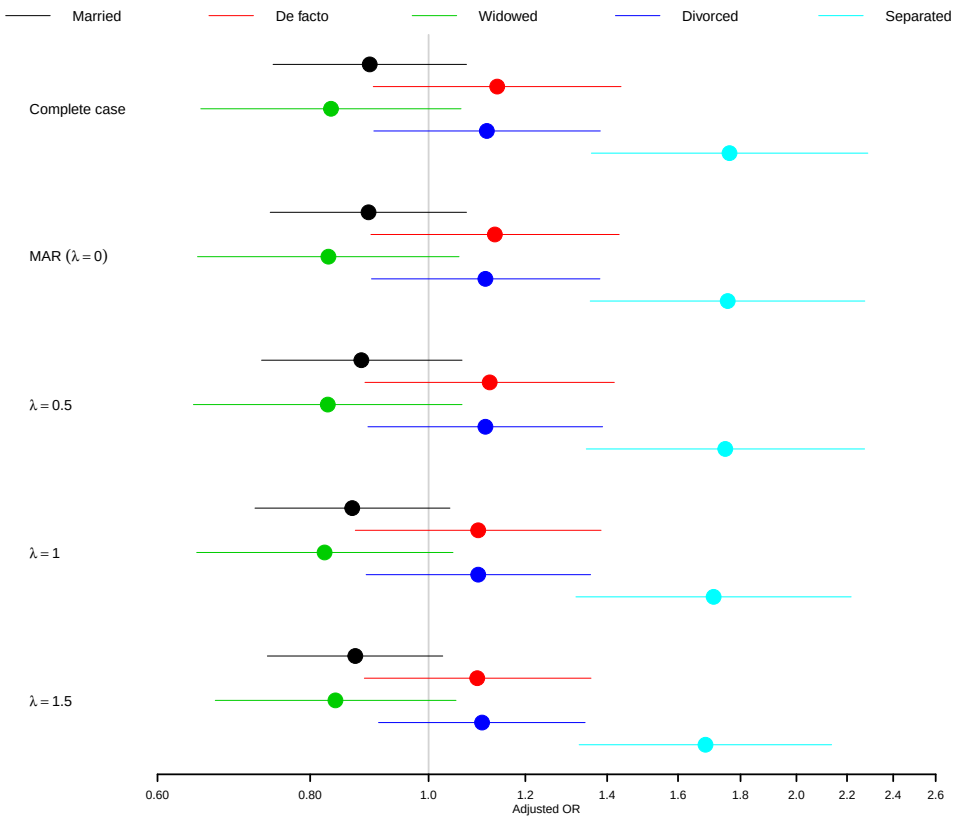

Figure 3: Estimated OR and 95% CI for marital status (Ref='Single'), under the complete case and selection model with different values of λ.

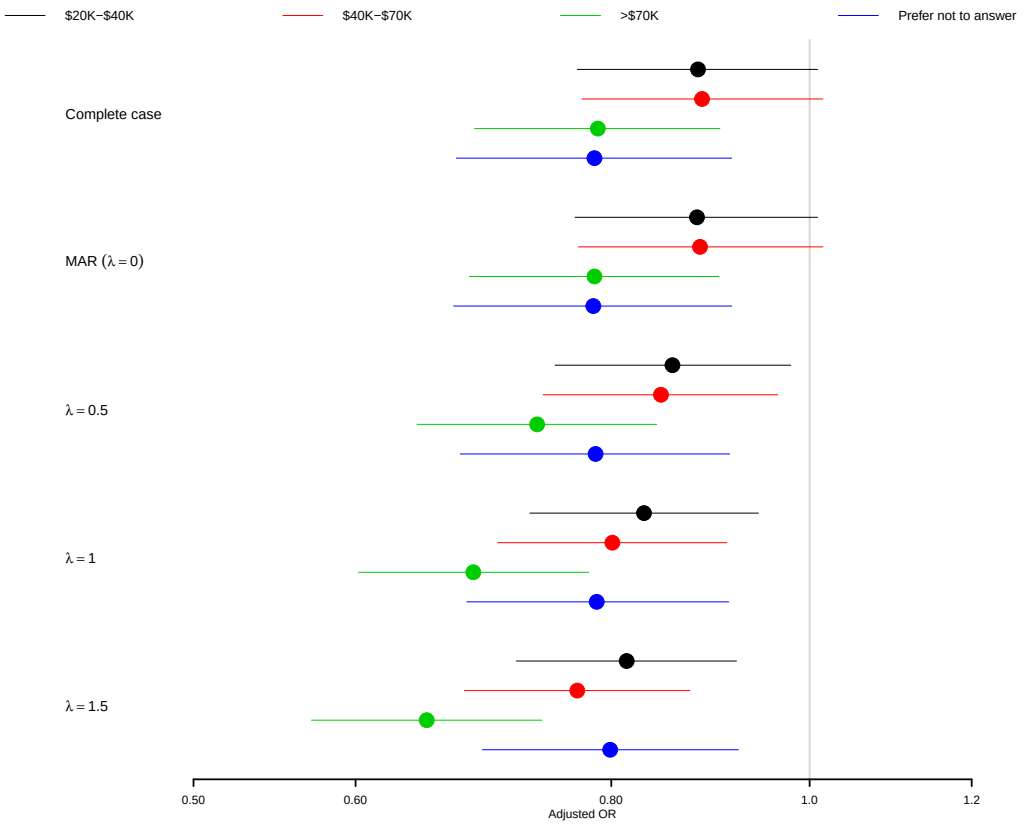

Figure 4: Estimated OR and 95% CI for household income (Ref='<$20,000'), under the complete case and selection model with different values of λ.

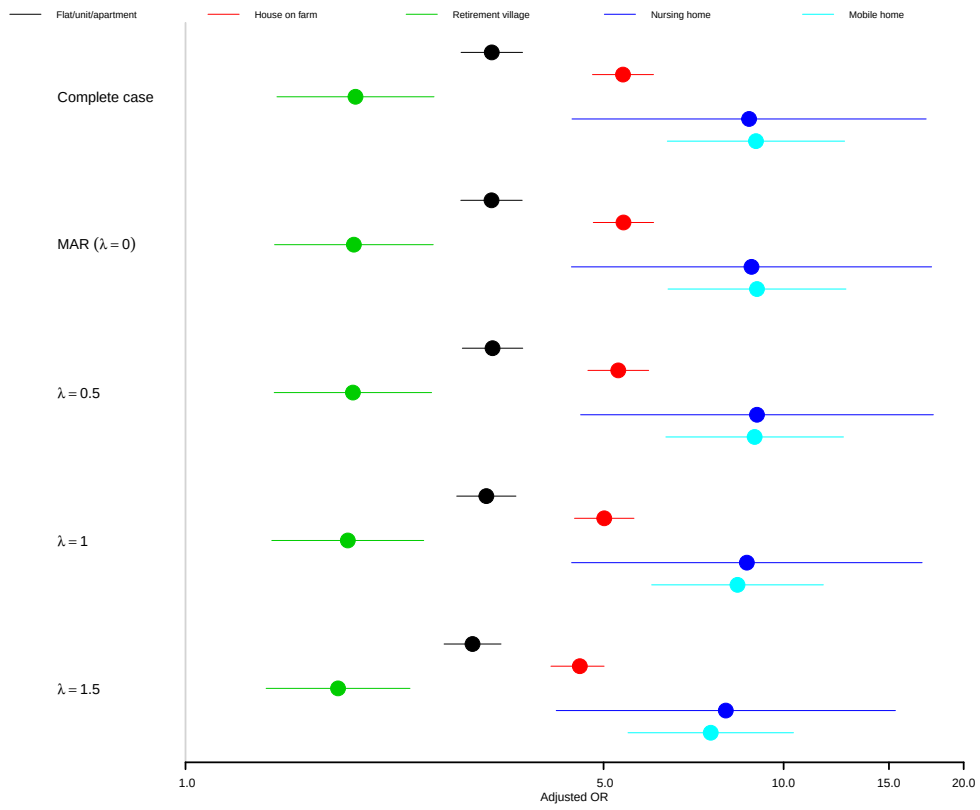

Figure 5: Estimated OR and 95% CI for dwelling-type (Ref='House'), under the complete case and selection model with different values of λ.

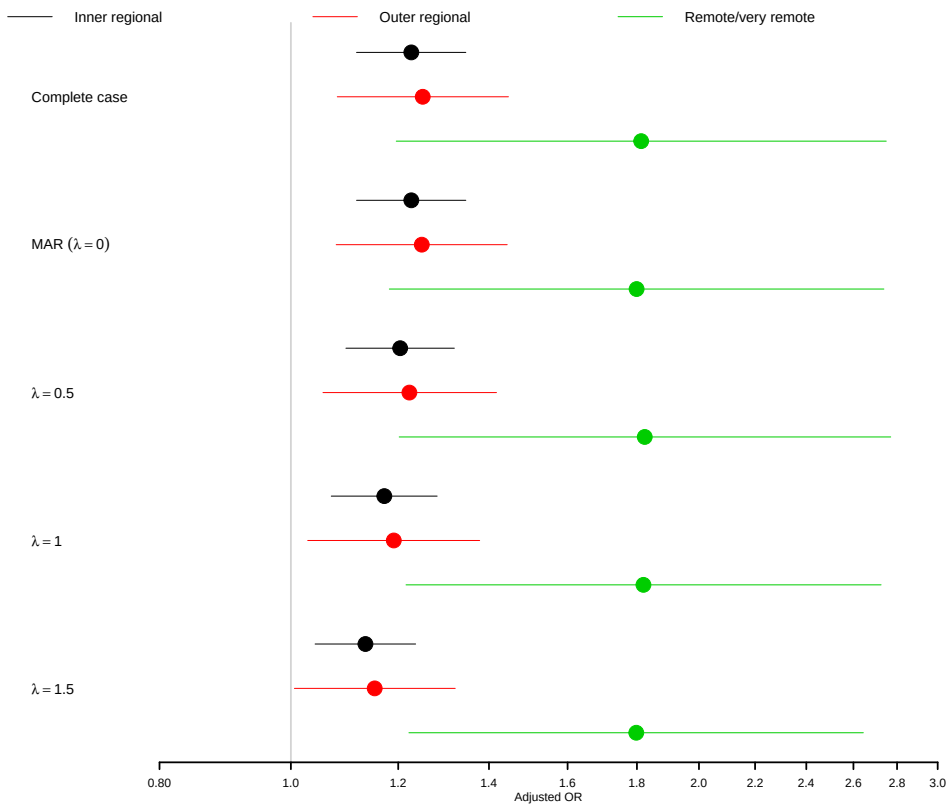

Figure 6: Estimated OR and 95% CI for area remoteness (Ref='Major cities'), under the complete case and selection model with different values of λ.

Appendix

Table A: Description of selected variables in the 45 and Up Study baseline survey

Baseline variable	Details
Gender	2 levels: 1=male, 2= female
Age group	5 levels: 45-54; 55-64; 65-74; 75-84; 85+
Country of birth	11 levels: Australia; North West Europe ('NW Europe'); Southern & Eastern Europe ('S & E Europe'); North Africa & Middle East ('Middle East'); South East Asia ('SE Asia'); North East Asia ('NE Asia'); Southern & Central Asia ('S & Central Asia'); America; Sub Saharan Africa; Oceania & Antarctica (not Australia) ('Oceania')
Speak language other than English at home	1=yes; 0=no
Marital status	6 levels: single; married; de facto; widowed; divorced; separated
Work status	7 levels: work full time or self-employed ('FT/self-employed'); work part time ('PT'); fully retired; partially retired; disabled/sick; look after home/study/unpaid work ('look after home'); unemployed
Household income	5 levels: $<$\$20,000; $\$20,000 - \$40,000$; $\$40,000 - \$60,000$; $\$60,000 - \$70,000$; $>$\$70,000
Dwelling-type	8 levels: house; flat/unit/apartment ('Flat/unit/apart.'); house on farm; retirement village; nursing home; hostel for the aged; mobile home; other
Carer status	1= carer; 0 = not a carer
Self-rated health	5 levels: excellent; very good; good; fair; poor
Physical function limitation (SF36)	5 levels: no function limitation; slight function limitation; moderate function limitation; significant function limitation; severe function limitation
Social connectedness	Duke Social Support Index (DSSI) subscale; divided into quartiles with higher levels representing worse social connectedness

REFERENCES

45 and Up Study Collaborators. Cohort profile: The 45 and Up Study. *International Journal of Epidemiology* 2008 Oct 37, 941-7.

Broadhead, W.E., Gehlbach, S.H., de Gruy, F.V. and Kaplan, B.H. (1988). The Duke-UNC Functional Social Support Questionnaire. Measurement of Social support in family medicine patients. *Medical Care*, 26(7):709-723.

Carpenter, J., Pocock, S., Lamm, C.J. (2002). Coping with missing data in clinical trials: a model based approach applied to asthma trials. *Statistics in Medicine*, 21:1043-1066.

Diggle, P. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43:49-93.

Etter, J.F and Perneger, T.V. (1997). Analysis of non-response bias in a mailed health survey. *Journal of Clinical Epidemiology*, 50:1123-1128.

Ibrahim J.G., Chen, M-H, Lipsitz S.R. (2001). Missing responses in generalized linear mixed models with then missing data mechanism is nonignorable. *Biometrika*, 88, 551-564.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

Mason, A., Best, N. Plewis, I. and Richardson, S. (2010). Insights into the use of Bayesian models for informative missing data. Technical report, Imperial College London.

Nohr, E.A., Frydenberg, M., Henriksen, T.B., Olsen, J. (2006) Does low participation in cohort studies induce bias? *Epidemiology*, 17(4):413-418.

Phongsavan, P., Grunseit, A.C., Bauman, A., Broom, D. Byles, J. Clarke, J. Redman, S. Nutbeam, D. and SEEF Project (2013). Age, gender, social contacts, and psychological distress: findings from the 45 and Up Study. *Journal of Aging Health*, 25(6):921-943.

RAND Health. (2009). Scoring Instructions for MOS 36-Item Short Form Survey Instrument (SF-36). [cited 2016; Available from:
`http://www.rand.org/health/surveys_tools/mos/mos_core_36item.html`

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387-394.

Scharfstein, D.O., Robins, J.M. and Rotnitzky, A. (1999). Adjustingfor nonignorable dropout using semiparametric nonresponse models (with discussion). *Journal of American Statistical Association*, 94:1096-1146.

Scharfstein, D.O., Daniels, M.J. and Robins J.M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, 4:495-512.

Uhrig, N.S.C. (2008). The nature and causes of attrition in the British Household Panel Survey. ISER Working Paper 2008-05. Colchester: Institute for Social and Economic Research, University of Essex.

Verbeke, G., Molenberghs, G., Thijes, H., Lesaffre, E. and Kenward, M.G. (2001) Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, 57:7-14.

Voorpostel, M. and Lipps, O. (2011). Attrition in the Swiss household panel: Is change associated with dropout? *Journal of Official Statistics* 27:301-318.

Watson, N. and Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In Methodology of longitudinal surveys, ed. Peter Lynn, 157?83. Chichester, UK: John Wiley & Sons.

Young, A.F., Powers, J.R., Bell, S.L. (2006). Attrition in longitudinal studies: who do you lose? *Australian and New Zealand Journal of Public Health*, 30(4):353-361.