Selection bias was reduced by recontacting non-participants

Juha Karvanen[1*], Hanna Tolonen[2], Tommi Härkänen[2], Pekka Jousilahti[2] and Kari Kuulasmaa[2]

[1] Department of Mathematics and Statistics, University of Jyvaskyla, Jyväskylä, Finland

[2] Department of Health, National Institute for Health and Welfare, Helsinki, Finland

February 29, 2016

Abstract

Objective: One of the main goals of health examination surveys (HES) is to provide unbiased estimates of health indicators at the population level. We demonstrate how multiple imputation methods may help to reduce the selection bias if partial data on some non-participants are collected.
Study Design and Setting: In the FINRISK 2007 study, a population-based health study conducted in Finland, a random sample of 10 000 men and women aged 25–74 were invited to participate. The study included a questionnaire data collection and a health examination. 6255 individuals participated in the study. Out of 3745 non-participants, 473 returned a simplified questionnaire after a recontact. Both the participants and the non-participants were followed up for death and hospitalizations. The follow-up data allowed to check the assumptions on the missing data mechanism and tailored multiple imputation methods were used to handle the missing data.
Results: Non-participation is a strong predictor for mortality in the five-year follow-up. However, the recontact response does not predict mortality or morbidity among the non-participants when adjusted for age and sex. The

result suggests that the recontact respondents can be used as proxy for all non-participants. A comparison of raw estimates and estimates adjusted for selection bias reveals clear differences in the estimated population prevalences of smoking and heavy alcohol usage.
Conclusion: All efforts to collect data on non-participants are likely to be useful even if the response rate for the recontact remains low. Statistical analysis of the recontact respondents provides an indication of the extent of the selection bias, even in studies where follow-up data are not available to check the assumptions.

Keywords: survey; non-response; bias; missing data; multiple imputation

What is new?:

- Register data on cause specific mortality and morbidity can be used to check the assumptions on the missing data mechanism.
- In the FINRISK 2007 study, we found that while the participants and the non-participants clearly differ, the non-participants with and without a recontact response have similar mortality and morbidity when adjusted for age and sex.
- We propose a multiple imputation approach to handle data missing not at random. The approach is applicable when data on a non-response questionnaire are available.

1 Introduction

Health examination surveys (HES) provide objective information about the health and health behaviors of the general population. Such data facilitate evidence-based policy decision and they can be used to plan and evaluate health promotion activities and for research.

Selective non-participation poses a major threat to the population representativeness of HESs. Especially when the aim is to monitor information about the target population, the representativeness of the results is essential [1, 2]. An estimate of a health indicator may suffer from selection bias if the decision of participating HES depends on a variable associated with the health indicator. The lower the participation rate, the more serious the problem might be [3].

The data on the participants alone offers no means to evaluate the impact of the non-participation. Sampling frames often contain information

on the age, sex and area of residence also for non-participants. Earlier studies have reported that young men and persons living in big cities are overrepresented among the non-participants [4–7]. Studies where the HES data have been linked with data from administrative registries covering also the non-participants reveal that there typically are significant differences between the participants and the non-participants. It has been found that non-participants have lower education, more often live on social welfare and are more often unemployed than participants [8]. Non-participants are reported to use more alcohol than participants [9, 10] and to have higher smoking and alcohol related mortality [11]. Non-participants of health surveys also more often use out-patient health care and have higher hospitalization rates than participants [4, 6, 7, 12, 13], have more psychotropic prescriptions [14], and have a higher mortality rate during the follow-up [15–17].

The sampling frame and administrative registries are not the only potential sources of data on the non-participants. The non-participants can be contacted again and asked to provide answers to a limited set of questions, so called non-response questionnaire. This kind of recontact data collection should not be mixed with reminders which are sent to make the person participate in the actual health examination. In the situation we consider, the sample can be divided into three non-overlapping participation groups:

1. the participants who both took part in the physical measurements and returned the questionnaire,

2. the non-participants of health examinations who after the recontact returned the non-response questionnaire and

3. the non-participants of health examinations who after the recontact did not return the non-response questionnaire.

Our objective is to provide unbiased estimators for health indicators, especially for the prevalence of daily smoking and heavy alcohol usage. A priori all three groups differ from each other and assumptions on the missing data mechanism are needed for unbiased estimation. We apply graphical models [18–20] to describe the assumptions on the missing data mechanism. The validity of the assumptions is evaluated using follow-up data on mortality and morbidity. The rationale is that any major differences in risk factors between the groups should reflect as a difference in total and cause specific mortality and morbidity. The missing data are handled with a multiple imputation (MI) method where the imputation model is built according to the assumed missing data mechanism.

2 Data and methods

2.1 Data

The FINRISK 2007 study is a cross-sectional population-based HES including a self-administered questionnaire, physical measurements such as the blood pressure and anthropometric measurements, and the collection of biological samples. The study was conducted in five areas: the Provinces of North Karelia and Kuopio in Eastern Finland, Turku-Loimaa region in Southwestern Finland, cities of Helsinki and Vantaa in Southern Finland and Oulu province in Northern Finland. A random sample of 10 000 men and women aged 25–74 years was selected from the National Population Register. The sampling was stratified by age group (10 year intervals), sex and geographical region.

The data available for everyone selected for the sample included background variables Z (age, sex and geographical region) as well as follow-up data on mortality and morbidity. To obtain mortality and morbidity data T, the whole FINRISK 2007 sample (both participants and non-participants) was linked to the National Causes of Death Register and the National Hospital Discharge Register using the personal identification code. Causes of death and reasons for the hospitalization, classified using ICD-10, were obtained until the end of 2012. For the analysis, we have used the grouping of the causes of death and hospitalizations presented in the Appendix A.

Each selected person received a letter of invitation together with the survey questionnaire. The invetees were instructed to fill in the questionnaire at home and return it during the health examination. The participation rate for the health examination was 63% and three persons explicitly refused any further contact. The same questionnaire was sent again to those who did not participate in the health examination after a reminder and did not explicitly refuse. The recipients were asked to return the filled questionnaire by mail. The response rate for the recontact was 13%. Selection variable M_1 has value 1 if the person participated in the health examination and 0 otherwise. Selection variable M_2 has value 1 if a non-participant returned the filled questionnaire and 0 otherwise.

The questionnaire data X are available for participation groups 1 and 2. The questionnaire included a large number of questions on lifestyle and health behavior. Questions on alcohol usage and smoking were used to derive indicators (binary variables) for heavy alcohol usage and daily smoking which are the main health indicators of interest in our analysis. Auxiliary variables used in the analysis include self-reported marital status, level of education (high, middle, low), self-reported hypertension, recency of blood pressure

measurement and self-reported high cholesterol.

The data on physical measurements Y are available only for the participants. These include measurements on body-mass index, systolic and diastolic blood pressure and total cholesterol.

The health indicators of interest include:

- the prevalence of heavy alcohol use (self-reported questionnaire data),
- the prevalence of daily smoking (self-reported questionnaire data),
- the prevalence of obesity (measured body-mass index ≥ 30 kg/m^2),
- the prevalence of high blood pressure (measured systolic blood pressure ≥ 140 mmHg), and
- the prevalence of elevated total cholesterol (measured total cholesterol ≥ 5 mmol/l).

2.2 Causal model

Causal model with design [18] for the FINRISK 2007 study is presented in Figure 1. The graphical model describes the assumed causal relationships together with the study design and the hypothesized missing data mechanism. The invited cohort $\{i : m_{1i} = 1\}$ is selected in random from the population strata defined by background variables Z_i. The decision of participation M_{1i} depends on background variables Z_i as well as variables X_i and Y_i to be measured in the questionnaire and in the health examination, respectively. Non-participants $\{i : M_{1i} = 0\}$ make another decision M_{2i} whether or not to provide questionnaire data after the recontact.

Physical measurements Y_i^* are available for the participants $\{i : M_{1i} = 1\}$. Measurements X_i^* on the questionnaire variables are available for the participants $\{i : M_{1i} = 1\}$ (participation group 1) and for the non-participants of the health examination who after recontact returned the non-response questionnaire $\{i : M_{1i} = 0, M_{2i} = 1\}$ (participation group 2). The questionnaire variables include self-reported health indicators as well as auxiliary variables. Measurements Z_i^* on the background variables and the dates T_i^* of the hospital discharges and deaths up to the end of 2012 are available for all individuals invited to the cohort $\{i : m_{1i} = 1\}$ (participation groups 1, 2 and 3).

Two assumptions called 'A' and 'B' are marked with dash-dotted edges in the graph. Assumption A concerns the decision returning the non-response questionnaire after the recontact. The dash-dotted edges $X_i \rightarrow M_{2i}$ and

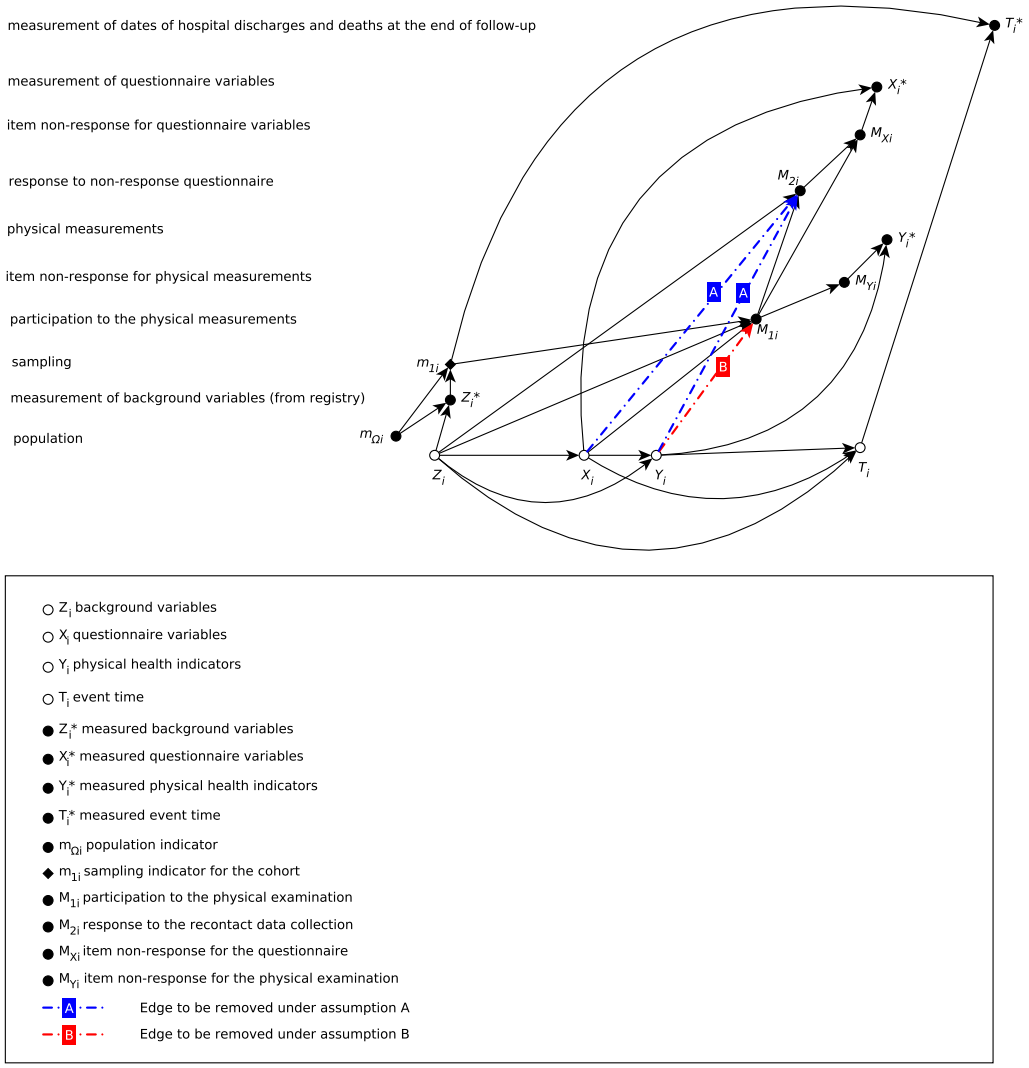

Figure 1: Causal model with design for the FINRISK 2007 study. The labels on the left of the graph describe the stages of the data collection and the locations of the nodes on the horizontal axis tell the causal ordering of the variables.

$Y_i \rightarrow M_{2i}$ marked with 'A' describe the effect of the survey variables on the decision responding in the recontact data collection. We study empirically whether these edges are necessary in the graph. If M_{2i} is independent on X_i and Y_i given background variables Z_i, the edges can be removed. This would imply $p(X_i \mid M_{2i} = 1, M_{1i} = 0, Z_i) = p(X_i \mid M_{2i} = 0, M_{1i} = 0, Z_i)$, which means that in the estimation of the averages of questionnaire variables, the participation group 2 can be used to represent all non-participants when conditioned on the background variables.

Assumption B concerns the impact of the physical health indicators on the decision participating in the health examination. In the graph, this impact is shown by the dash-dotted edge $Y_i \rightarrow M_{1i}$ marked with 'B'. If it is assumed that this edge does not exist, i.e. the HES participation does not depend on the physical health indicators given the background variables and the questionnaire variables, the conditional distribution $p(Y_i \mid X_i, Z_i)$ can be estimated from the data on the participants. This conditional distribution can then be used together with the distribution $p(X_i, Z_i \mid M_{1i} = 0)$ estimated from the data on the recontact respondents to obtain an estimate for the marginal distribution of Y_i for the non-participants. Differently from the assumption A, the validity of assumption B cannot be studied empirically with our current data.

Causal models with design for some alternative scenarios are presented in Appendix B.

2.3 Statistical methods for assumption checking

To check the validity of assumption A, we fit a statistical model that explains the risk of death and hospitalization by selection variables M_1 and M_2. Background variables Z are included as covariates. If the regression coefficients for M_1 and M_2 differ from zero and the differences are statistically and practically significant, we conclude that the risks differ between the participation groups. It follows that the distribution of some risk factors must also differ between the participation groups. In the opposite case, we conclude that the risk factors do not differ between the participation groups when adjusted for the background variables. In particular, assumption A is validated if the regression coefficient for M_2 is close is zero. Theoretically, there exists a possibility that the differences in one risk factor are incidentally canceled by the differences in other risk factors such that the risks are the same in all the groups but this is considered unlikely in the practice.

All statistical analyses are carried out with R [21]. Deaths and hospitalizations are analysed separately. Logistic regression is used to model the death during the follow-up and the sampling weights are applied. Two mod-

els are fitted: a model for all individuals in the sample and a separate model for non-participants $\{i : M_{1i} = 0\}$ only. The predictors common for both models are sex, age and region. Both models also include recontact response indicator M_2 as a predictor. In addition, the model for all individuals includes participation indicator M_1 as a predictor.

Zero-inflated negative binomial regression model is used to model the number of the hospital visits during the follow-up. The predictors are the same as in the models for the deaths. R package `pscl` [22] is utilized.

2.4 Statistical methods for estimation of health indicators

In the estimation of the population averages of health indicators, MI [23, 24] is used to handle the missing values of questionnaire variables X and physical measurement variables Y. In addition to missing data due to non-participation, there are occasional item non-response in questionnaire variables. The imputation method is called MI-MNAR where MNAR tells that the data are assumed to be missing not at random.

The possible bias of Rubin's MI variance estimator for data that are collected with a complex sample design has been discussed by many authors [25–28]. Although unbiased estimation cannot be guaranteed theoretically, MI methods seem to work well in the practice [25, 29]. In our analysis, the imputation model includes the stratification variables age, sex and region as explanatory variables. Interactions between sex and age and between sex and region are included. The sampling weights are applied when estimates and their variances are calculated from the imputed datasets.

The questionnaire variables daily smoking, heavy alcohol usage, marital status, level of education, self-reported hypertension, recency of blood pressure measurement and self-reported high cholesterol are imputed variable-by-variable using fully conditional specification [24]. These variables added with age, sex and region are used as explanatory variables in the imputation models for each other. Interactions between sex and age and between sex and region are included in all imputation models. Assumption A is implemented by using participation indicator M_1 as an interaction term for the explanatory variables. This means that only the data from participation group 2 are used to estimate the imputation model for the missing questionnaire variables in participation group 3.

The physical measurement variables are imputed as follows:

- obesity is explained by sex, age, region, level of education, marital status, heavy alcohol usage and daily smoking,

- high blood pressure is explained by sex, age, region, level of education, marital status, heavy alcohol usage, daily smoking, obesity, recency of blood pressure measurement and self-reported hypertension,

- elevated total cholesterol is explained by sex, age, region, level of education, marital status, heavy alcohol usage, daily smoking, obesity and self-reported high cholesterol.

From assumption B it follows that indicators M_1 and M_2 are not needed here. Interactions between sex and age and between sex and region are included in all imputation models.

The imputation model is a logistic regression model for binary variables and predictive mean matching for all other variables. R package `mice` [30] is utilized and 50 imputations are generated.

2.5 Alternative statistical methods for estimation of health indicators

For methodological comparisons, two alternative imputation methods called MI-MAR and MI-MAR (no recontact) are also applied. MI-MAR is an imputation method similar to MI-MNAR with an exception that participation indicator M_1 is not used as an interaction term in the imputation models. This means that the data are assumed to be missing at random and the data from both participation groups 1 and 2 are used to estimate the imputation model for the missing questionnaire variables in participation group 3. The comparison between MI-MNAR and MI-MAR provides an indication of the extent of selection bias that cannot be removed only by conditioning on the background variables Z.

MI-MAR (no recontact) uses otherwise the same imputation method as MI-MAR but simulates the situation where non-participants are not recontacted. The comparison between MI-MNAR and MI-MAR provides an indication of the importance of the recontact.

To compare MI-MNAR with the MI-MAR methods, we also impute the deaths and hospital visits during the follow-up. As these are available for the full cohort, the imputation based estimates can be benchmarked against the real data [29, 31]. In the imputation, it is assumed that the deaths and hospital visits are available for participation groups 1 and 2 but missing for participation group 3 and the imputation model is similar to the model used for the questionnaire variables.

3 Results

The basic demographics of the three participation groups are shown in Table 1. The non-participants without a recontact response have a lower mean age and a higher proportion of men than the participants. However, the age and sex distributions of the non-participants with the recontact response seem to resemble the participants rather than the non-participants without recontact response.

Logistic regression models for the death during the follow-up are presented in Table 2. In both models, the point estimate of the regression coefficient for the recontact response is practically zero, which means that the variable is not a predictor of death. On the contrary, participation to the physical measurements is a very strong predictor of death. The model indicates that the difference between the participants and the non-participants is equivalent to the age difference of 12.5 years. Region is not included in the final model because it did not predict death during the follow-up.

The zero-inflated negative binomial regression model for the number of hospital visits (excluding pregnancy related visits with ICD-10 codes O00-O99) is presented in Table 3. The results show that the non-participants have more hospital visits than the participants but the differences between non-participants with and without a recontact response are very small. The results shown in Tables 2 and 3 provide the empirical justification for the assumption A in the causal model in Figure 1.

Table 4 provides a summary of the causes of death and the causes of hospitalization in the three participation groups. It can be seen that the total mortality rate for the non-participants in 5 year follow-up is more than twice the total mortality rate for the participants. Even larger differences are seen for specific causes of death. The ratio of the mortality rates for the non-participants versus the participants is over three for alcohol related causes and about four for smoking related causes. The results support the interpretation that heavy alcohol usage and smoking are more common among the non-participants than among the participants.

The non-participants also have more hospital visits than participants. The difference is clear in hospital visits due to alcohol related causes but there are differences also in hospital visits due to infections and smoking related causes.

Table 1: Basic demographics for the participation groups. All proportions are standardized to the population level using the sampling weights determined by age, sex and region. 95% confidence intervals are presented in the parentheses.

	Participants	Non-participants with recontact response	Non-participants without recontact response
N	6257	473	3270
Women, %	54.1 (52.6,55.6)	54.7 (49.2,60.2)	44.0 (41.9,46.1)
Mean age, years	48.9 (48.6,49.2)	47.3 (46.0,48.5)	44.6 (44.2,45.1)
Education			
High, %	37.4 (36.0,38.8)	28.0 (23.0,32.9)	—
Low, %	28.9 (27.6,30.2)	35.8 (30.5,41.1)	—
Daily smokers, %	21.8 (20.5,23.0)	33.4 (28.2,38.6)	—
Heavy alcohol users, %	5.2 (4.6,5.9)	6.4 (3.7,9.1)	—
Self-reported hypertension,%	59.8 (58.3,61.2)	56.6 (51.1,62.1)	—
Self-reported high cholesterol, %	45.1 (43.6,46.6)	40.7 (35.3,46.1)	—
Civil status			
Married, %	54.0 (52.5,55.4)	46.8 (41.3,52.3)	—
Cohabiting, %	16.9 (15.8,18.0)	18.4 (14.1,22.7)	—
Single, %	15.5 (14.4,16.6)	20.7 (16.2,25.1)	—
Divorced, %	10.4 (9.5,11.3)	11.2 (7.7,14.7)	—
Widow, %	3.2 (2.7,3.7)	2.8 (1.0,4.7)	—
Time from the last cholesterol measurement			
Less than half year, %	20.7 (19.5,21.9)	25.4 (20.6,30.2)	—
Half year to one year, %	17.5 (16.3,18.6)	17.1 (13.0,21.3)	—
One to five years, %	31.2 (29.8,32.5)	24.3 (19.6,29.1)	—
Over five years, %	10.8 (9.9,11.7)	8.8 (5.7,12.0)	—
Never measured, %	12.9 (11.9,13.9)	16.1 (12.1,20.2)	—
Do not know, %	7.0 (6.2,7.7)	8.2 (5.2,11.2)	—
Time from the last blood pressure measurement			
Less than half year, %	49.7 (48.2,51.1)	53.4 (47.9,58.9)	—
Half year to one year, %	20.2 (19.1,21.4)	21.0 (16.5,25.5)	—
One to five years, %	25.2 (23.9,26.4)	21.2 (16.7,25.7)	—
Over five years, %	4.3 (3.7,4.9)	2.4 (0.7,4.0)	—
Never measured, %	0.7 (0.4,0.9)	2.0 (0.4,3.5)	—

Table 2: Estimated parameters with standard errors (SE) from the logistic regression models for the death during the follow-up.

Predictor	All data Estimate (SE)	Participation groups 2 and 3 Estimate (SE)
Intercept	-6.50 (0.35)	-6.67 (0.43)
Age (10 years)	0.92 (0.05)	0.94 (0.07)
Sex (Woman)	-0.88 (0.12)	-0.86 (0.16)
Participant (Yes)	-1.15 (0.12)	—
Recontact respondent (Yes)	-0.07 (0.22)	-0.08 (0.23)

Table 3: Estimated parameters with standard errors (SE) from the zero-inflated negative binomial regression model for the number of hospital visits.

Count model	Estimate (SE)
Intercept	-1.54 (0.33)
Age: Men (10 years)	0.41 (0.03)
Age: Women (10 years)	0.34 (0.02)
Sex (Woman)	0.24 (0.20)
Region: Kuopio	0.04 (0.08)
Region: Turku/Loimaa	-0.16 (0.08)
Region: Helsinki/Vantaa	-0.46 (0.07)
Region: Oulu	0.05 (0.07)
Participant (Yes)	-0.54 (0.06)
Recontact respondent (Yes)	-0.17 (0.11)
Zero model	Estimate (SE)
Intercept	1.82 (0.77)
Age (10 years)	-0.50 (0.10)
Sex (Women)	-0.81 (0.42)
Participant (Yes)	-1.32 (0.61)
Recontact respondent (Yes)	-0.01 (0.49)

Table 4: Mortality, morbidity, causes of death and causes of hospitalization by the participation group. Standardization (by age, sex and region) is done with respect to the background population.

	Participants	Non-participants with recontact response	Non-participants without recontact response
N	6257	473	3270
Number of deaths	166	34	204
Deaths per 1000 (95% CI)	26.5 (22.5,30.5)	71.9 (48.6,95.2)	62.4 (54.1,70.7)
standardized (95% CI)	22.1 (18.5,25.7)	55.1 (34.6,75.7)	49.5 (42.1,57.0)
Hospital visits	7803	817	5031
Hospital visits per 1000 (95% CI)	1247 (1175,1319)	1727 (1370,2084)	1539 (1405,1673)
standardized (95% CI)	1052 (989,1115)	1357 (1043,1670)	1316 (1196,1437)
Cause of death, standardized number of cases per 1000 (absolute number of cases)			
Cardiovascular diseases	7.0 (56)	18.0 (11)	13.2 (58)
Cancer	8.0 (62)	18.2 (8)	9.9 (48)
Infections	0.7 (4)	0	0.3 (2)
Injuries, poisonings and external causes	2.6 (14)	3.7 (3)	8.6 (23)
Suicide	1.3 (8)	3.7 (3)	4.0 (9)
Other	3.9 (30)	15.2 (12)	17.4 (73)
Alcohol related	1.5 (7)	4.3 (3)	5.1 (20)
Smoking related	0.7 (8)	5.2 (2)	2.9 (15)
Cause of hospitalization, standardized number of visits per 1000 (absolute number of visits)			
Cardiovascular diseases	150 (1216)	173 (134)	160 (663)
Cancer	105 (845)	123 (61)	68 (317)
Infections	52 (405)	150 (78)	79 (340)
Injuries, poisonings and external causes	119 (756)	193 (96)	164 (513)
Other	637 (4581)	787 (448)	861 (3198)
Alcohol related	11 (81)	72 (45)	72 (248)
Smoking related	6 (63)	29 (11)	12 (42)

Table 5 presents the health indicators estimated using the proposed MI-MNAR method for the missing data. The MNAR estimates of the population prevalences of heavy alcohol usage and smoking clearly differ from the estimates from the participants only. It can also be seen that simpler MI methods provide estimates that clearly differ from the MI-MNAR estimates. The estimated population prevalences of overweight (BMI $\geq$ 30 kg/m^2) , high blood pressure (systolic blood pressure $\geq$ 140 mmHg) and high cholesterol (total cholesterol $\geq$ 5.0 mmol/l) are practically same for all estimation methods and close to the estimates calculated from the participants only.

The last two columns of Table 5 provide the estimates for the number of deaths and the number of hospital visits. For these statistics, benchmark numbers from the full cohort are available. It can be seen that the point estimates from the full cohort locate outside the confidence intervals calculated by the MI-MAR methods. The same holds true also for the confidence intervals obtained from participants only. The MI-MNAR estimates are higher than the full cohort estimates but the MI-MNAR confidence intervals nevertheless contain the full cohort point estimates. The selection bias increases uncertainty, which leads to wider confidence intervals in the MI-MNAR approach.

Table 5: The estimates of the health indicators for the different participation groups and the population level estimates with three methods of MI. MI-MNAR is the proposed MI method and MI-MAR with and without recontact are simpler MI methods that suffer from the selection bias.

	Heavy alcohol users % (95% CI)	Daily smokers % (95% CI)	BMI $\geq$ 30 kg/m^2 % (95% CI)	Systolic blood pressure $\geq$ 140 mmHg % (95% CI)	Total cholesterol $\geq$ 5.0 mmol/l % (95% CI)	Deaths per 1000 (95% CI)	Hospitalizations per 1000 (95% CI)
Full cohort	—	—	—	—	—	33.0 (29.5,36.5)	1157 (1098,1215)
Participants	5.2 (4.7,5.8)	21.8 (20.7,22.8)	21.0 (20.0,22.0)	30.7 (29.5,31.8)	58.7 (57.4,59.9)	22.1 (18.5,25.7)	1052 (989,1115)
Recontact respondents	6.4 (4.1,8.7)	33.4 (29.1,37.7)	—	—	—	55.1 (34.6,75.7)	1357 (1043,1670)
MI-MNAR	6.8 (5.6,8.1)	27.1 (24.8,29.3)	20.8 (19.7,22.0)	29.3 (28.1,30.6)	58.0 (56.5,59.5)	39.8 (32.7,47.0)	1303 (1175,1431)
MI-MAR	5.5 (4.8,6.3)	23.4 (22.3,24.6)	20.4 (19.3,21.6)	29.1 (27.9,30.3)	57.5 (56.2,58.9)	23.6 (19.7,27.5)	1027 (968,1085)
MI-MAR, no recontact	5.4 (4.7,6.0)	23.4 (22.3,24.6)	20.1 (19.0,21.3)	29.1 (27.9,30.2)	57.7 (56.2,59.1)	23.4 (19.7,27.1)	1030 (966,1095)

4 Discussion

We have studied the methods for reducing selection bias with data from FINRISK 2007. In addition to questionnaire data and physical measurements data for the participants, the data included also follow-up data for both participants and non-participants and questionnaire data for the recontact respondents. These data components allowed us to develop a MI method tailored for the current study. The follow-up data, which will not be available for timely survey reporting, was used only for checking the assumptions on the missing data mechanism.

As expected, we found that non-participation is a strong predictor for mortality in the five-year follow-up. This result is consistent with other studies [15–17] and indicates that the data are MNAR. However, rather surprisingly, the recontact response did not predict mortality or morbidity among the non-participants when adjusted for age, sex and region. The result was utilized in the MI-MNAR method devised to reduce selection bias. The comparison of raw estimates and estimates adjusted for selection bias reveals clear differences in the estimated population prevalences of smoking and heavy alcohol usage.

The validity of the results depends on the validity of assumptions A and B concerning the non-response mechanism. The follow-up data supported assumption A but the FINRISK 2007 data could not be used to study the validity of assumption B. Thus, the prevalence estimates for heavy alcohol usage and daily smoking can be considered more plausible than the prevalence estimates for obesity, hypertension and elevated total cholesterol because the former estimates require only assumption A to hold whereas the latter estimates require both assumptions A and B to hold.

The strength of the proposed MI-MNAR approach lies in its ability to utilize the data on the recontact respondents according to assumption A. On the contrary, the MI-MAR method used a stronger assumption on the similarity of participants and non-participants conditioning on the background variables and failed to produce credible estimates for the prevalences of smoking and heavy alcohol usage. In addition, the MI-MAR approach underestimates the uncertainty in the estimates. We conclude that methods based on the MAR assumption for participation were insufficient for removing the selection bias and should not be used here. If assumption A was not valid or could not be verified, alternative models, such as repeated attempt selection model [32] or repeated attempt pattern mixture model [33], could have been applied.

MI was used for the missing data problem but it is not the only option available. An alternative approach would have been a combination of MI

and inverse probability weighting (IPW) [25]. In this approach, MI is first applied to data containing participation groups 1 and 2. Individuals in these participation groups are then re-weighed to also represent the persons in participation group 3. Bayesian methods could have been used as well.

From our current study, we can conclude the following three main messages. First, non-participation is a serious problem in HESs. In FINRISK 2007, this is seen by comparing the mortality and morbidity rates between the participants and non-participants as well as comparing the estimates for population level health indicators with and without correcting for selective non-participation. Second, the follow-up data for the whole sample is useful in checking the essential assumptions on the missing data mechanism. In FINRISK 2007, the follow-up data revealed that there must be large differences in the health status and the risk factors between the participants and the non-participants. The follow-up data also revealed that the decision responding in the recontact data collection does not seem to depend on the physical measurements or questionnaire variables. Third, efforts to collect data on non-participants may turn out to be worthwhile even if the response rate for the recontact remains low. In FINRISK 2007, only 13% of the non-participants responded in the recontact data collection but these data have a central role when the health indicators are estimated for the whole sample.

Further analyses are needed to check if our empirical finding on the non-participants with and without a recontact response holds also in other cohorts from Finland and elsewhere. For instance, the Leiden 85-plus Study ($n = 692$) reached a conclusion that the non-participants with and without a recontact response differ by their health status [34].

Our results showed marked changes in the health indicators for heavy alcohol use and daily smoking when adjusted using the MI-MNAR method. On the other hand, only minor changes were observed for obesity, hypertension and elevated total cholesterol. This reflects the importance of additional information on the non-participants. The non-response questionnaire had specific questions on alcohol use and daily smoking, while there were no good proxy questions for obesity, hypertension and elevated total cholesterol to be used for non-participants. For obesity, we could have obtained better estimates if we had had self-reported height and weight for those responding to the non-response questionnaire. Unfortunately, this information was not asked.

Partial questionnaire information from some of the health examination survey non-participants does not remove the non-participation bias completely but it provides valuable information which can be used to estimate the magnitude of the non-participation bias. There is some evidence indicating that the prevalences of health behaviors, such as smoking [9, 35] and

alcohol use [9, 10, 36], and the prevalences of health conditions, such as mental disorders [9, 37] and obesity [38, 39], would be underestimated if based on survey participants only. Even in studies where follow-up data are not available to check the assumptions, the statistical analysis of the recontact respondents provides an indication of the extent of the selection bias.

Representative estimates for the levels of risk factors at the population are needed for the health policy decision making and the planning and evaluation of prevention programs. Therefore, any additional information which helps us to improve our population level estimates is needed.

Acknowledgements

This work was supported by Academy of Finland [grant number 266251].

References

[1] Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. International Journal of Epidemiology. 2013;42(4):1012–1014.

[2] Ebrahim S, Smith GD. Commentary: Should we always deliberately be non-representative? International Journal of Epidemiology. 2013;42(4):1022–1026.

[3] Nishimura R, Wagner J, Elliott M. Alternative Indicators for the Risk of Non-response Bias: A Simulation Study. International Statistical Review. 2015;DOI: 10.1111/insr.12100.

[4] Uusküla A, Kals M, McNutt LA. Assessing non-response to a mailed health survey including self-collection of biological material. The European Journal of Public Health. 2011;21:538–542.

[5] Søgaard AJ, Selmer R, Bjertness E, Thelle D. The Oslo Health Study: The impact of self-selection in a large, population-based survey. International Journal for Equity in Health. 2004;3(1):3.

[6] Laaksonen M, Aittomäki A, Lallukka T, Rahkonen O, Saastamoinen P, Silventoinen K, et al. Register-based study among employees showed small nonparticipation bias in health surveys and check-ups. Journal of Clinical Epidemiology. 2008;61(9):900–906.

[7] Kjøller M, Thoning H. Characteristics of non-response in the Danish Health Interview Surveys, 1987–1994. The European Journal of Public Health. 2005;15(5):528–535.

[8] Drivsholm T, Eplov LF, Davidsen M, Jørgensen T, Ibsen H, Hollnagel H, et al. Representativeness in population-based studies: a detailed description of non-response in a Danish cohort study. Scandinavian Journal of Public Health. 2006;34(6):623–631.

[9] Torvik FA, Rognmo K, Tambs K. Alcohol use and mental distress as predictors of non-response in a general population health survey: the HUNT study. Social Psychiatry and Psychiatric Epidemiology. 2012;47(5):805–816.

[10] Gray L, McCartney G, White IR, Katikireddi SV, Rutherford L, Gorman E, et al. Use of record-linkage to handle non-response and improve alcohol consumption estimates in health survey data: a study protocol. BMJ open. 2013;3(3):e002647.

[11] Christensen AI, Ekholm O, Gray L, Glümer C, Juel K. What is wrong with non-respondents? Alcohol-, drug-and smoking related mortality and morbidity in a 12-year follow up study of respondents and non-respondents in the Danish Health and Morbidity Survey. Addiction. 2015;.

[12] Osler M, Schroll M. Differences between participants and non-participants in a population study on nutrition and health in the elderly. European Journal of Clinical Nutrition. 1992;46(4):289–295.

[13] Alkerwi A, Sauvageot N, Couffignal S, Albert A, Lair ML, Guillaume M. Comparison of participants and non-participants to the ORISCAV-LUX population-based study on cardiovascular risk factors in Luxembourg. BMC Medical Research Methodology. 2010;10(1):80.

[14] Vercambre MN, Gilbert F. Respondents in an epidemiologic survey had fewer psychotropic prescriptions than nonrespondents: an insight into health-related selection bias using routine health insurance data. Journal of Clinical Epidemiology. 2012;65(11):1181–1189.

[15] Thygesen LC, Johansen C, Keiding N, Giovannucci E, Grønbæk M. Effects of sample attrition in a longitudinal study of the association between alcohol intake and all-cause mortality. Addiction. 2008;103(7):1149–1159.

[16] Larsen SB, Dalton SO, Schüz J, Christensen J, Overvad K, Tjønneland A, et al. Mortality among participants and non-participants in a prospective cohort study. European Journal of Epidemiology. 2012;27(11):837–845.

[17] Jousilahti P, Salomaa V, Kuulasmaa K, Niemelä M, Vartiainen E. Total and cause specific mortality among participants and non-participants of population based health surveys: a comprehensive follow up of 54 372 Finnish men and women. Journal of Epidemiology and Community Health. 2005;59(4):310–315.

[18] Karvanen J. Study design in causal models. Scandinavian Journal of Statistics. 2015;42(2):361–377.

[19] Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Statistical Methods in Medical Research. 2012;21(3):243–256.

[20] Thoemmes F, Mohan K. Graphical representation of missing data problems. Structural Equation Modeling: A Multidisciplinary Journal. 2015;(ahead-of-print):1–13.

[21] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: `http://www.R-project.org/`.

[22] Zeileis A, Kleiber C, Jackman S. Regression Models for Count Data in R. Journal of Statistical Software. 2008;27(8).

[23] Rubin DB. Multiple imputation for nonresponse in surveys. Wiley Series in probability and mathematical statistics. New York: John Wiley & Sons, Inc.; 1987.

[24] Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: CRC Press; 2012.

[25] Seaman SR, White IR, Copas AJ, Li L. Combining Multiple Imputation and Inverse-Probability Weighting. Biometrics. 2012;68(1):129–137.

[26] Kim JK, Michael Brick J, Fuller WA, Kalton G. On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(3):509–521.

[27] Robins JM, Wang N. Inference for imputation estimators. Biometrika. 2000;87(1):113–124.

[28] Binder DA, Sun W. Frequency valid multiple imputation for surveys with a complex design. In: Proceedings of the Survey Methods Section. American Statistical Association; 1996. p. 281–286.

[29] Härkänen T, Karvanen J, Tolonen H, Lehtonen R, Djerf K, Juntunen T, et al. Systematic handling of missing data in complex study designs - experiences from the Health 2000 and 2011 Surveys. Journal of Applied Statistics. 2016;Accepted for publication.

[30] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1–67. Available from: `http://www.jstatsoft.org/v45/i03/`.

[31] Santin G, Geoffroy B, Bénézet L, Delézire P, Chatelot J, Sitta R, et al. In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse. Journal of Clinical Epidemiology. 2014;67(6):722–730.

[32] Alho JM. Adjusting for nonresponse bias using logistic regression. Biometrika. 1990;77(3):617–624.

[33] Daniels MJ, Jackson D, Feng W, White IR. Pattern mixture models for the analysis of repeated attempt designs. Biometrics. 2015;DOI: 10.1111/biom.12353.

[34] Bootsma-Van Der Wiel A, Van Exel E, De Craen A, Gussekloo J, Lagaay A, Knook D, et al. A high response is not essential to prevent selection bias: results from the Leiden 85-plus study. Journal of Clinical Epidemiology. 2002;55(11):1119–1125.

[35] Tolonen H, Dobson A, Kulathinal S, et al. Effect on trend estimates of the difference between survey respondents and non-respondents: results from 27 populations in the WHO MONICA Project. European Journal of Epidemiology. 2005;20(11):887–898.

[36] Lahaut VM, Jansen HA, van de Mheen D, Garretsen HF, Verdurmen JE, Van Dijk A. Estimating non-response bias in a survey on alcohol consumption: comparison of response waves. Alcohol and Alcoholism. 2003;38(2):128–134.

[37] Lundberg I, Thakker KD, Hällström T, Forsell Y. Determinants of non-participation, and the effects of non-participation on potential cause-effect relationships, in the PART study on mental disorders. Social Psychiatry and Psychiatric Epidemiology. 2005;40(6):475–483.

[38] Voigt LF, Koepsell TD, Daling JR. Characteristics of telephone survey respondents according to willingness to participate. American Journal of Epidemiology. 2003;157(1):66–73.

[39] Van Loon AJM, Tijhuis M, Picavet HSJ, Surtees PG, Ormel J. Survey non-response in the Netherlands: effects on prevalence estimates and associations. Annals of Epidemiology. 2003;13(2):105–110.

Appendix A: Classification of diseases and causes of death by ICD-10 codes

Disease/cause of death	ICD-10 codes
Cardiovascular diseases	I00–I99
Cancer	C00–C99
Infections	A00–A99, J22
Injuries, poisonings and external causes	S00–S99, T00–T99, U00–U99, V00–V99, X00–X99, Y00–Y99
Suicide	X60–X84
Alcohol related	G31.2, G62.1, G72.1, K29.2, K85.2, K86.0, I42.6, F10, K70, T51, X45, Y15
Smoking related	J41–J44, C34

Appendix B: Causal models for alternative scenarios

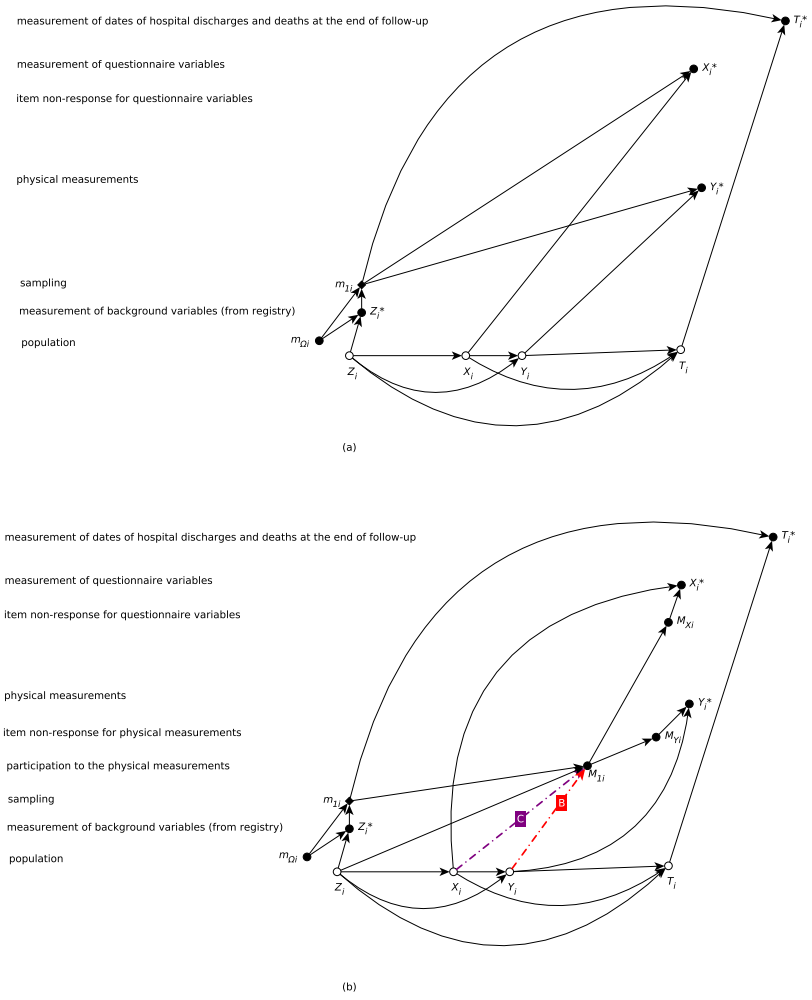

Figure B.1: Causal model with design for the FINRISK 2007 study in alternative scenarios. Panel (a) represents a scenario where there are no missing data. Panel (b) represents a scenario where the non-participants are not recontacted. Under the MAR assumption, the edges marked with B and C are removed.