Systematic handling of missing data in complex study designs - experiences from the Health 2000 and 2011 Surveys

Tommi Härkänen[a,*] Juha Karvanen[b†] Hanna Tolonen[c‡]
Risto Lehtonen[d§] Kari Djerf[e¶] Teppo Juntunen[f‖]
Seppo Koskinen[g**]

January 20, 2016

Abstract

We present a systematic approach to the practical and comprehensive handling of missing data motivated by our experiences of analyzing longitudinal survey data. We consider the Health 2000 and 2011 Surveys (BRIF8901) where increased non-response and non-participation from 2000 to 2011 was a major issue. The model assumptions involved in the complex sampling design, repeated measurements design, non-participation mechanisms and associations are presented graphically using methodology previously defined as a causal model with design i.e. a functional causal model extended with the study design. This tool forces the statistician to make the study design and the missing data mechanism explicit. Using the systematic approach, the sampling probabilities and the participation probabilities can be considered separately. This is beneficial when the performance of missing data methods are to be compared. Using data from Health 2000 and 2011 Surveys and from national registries, it was found that multiple imputation removed almost all differences between full sample and estimated prevalences. The inverse probability weighting removed more than half and the doubly robust method

[*a,*]Corresponding author; National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: Tommi.Harkanen@thl.fi, Tel.: +358295248719

[†b]University of Jyväskylä, P.O.Box 35, FI-40014 Jyväskylä, Finland, Email: juha.karvanen@jyu.fi, Tel.: +358295248719

[‡c]National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: hanna.tolonen@thl.fi, Tel.: +358295248638

[§d]University of Helsinki, P.O.Box 68, FI-00014 Helsinki, Finland, Email: risto.lehtonen@helsinki.fi, Tel.: +358919151404

[¶e]Statistics Finland, Työpajankatu 13, FI-00580 Helsinki, Finland, Email: kari.djerf@stat.fi, Tel.: +358917343425

[‖f]National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: teppo.juntunen@thl.fi, Tel.: +358295247215

[**g]National Institute for Health and Welfare, P.O.Box 30, FI-00271 Helsinki, Finland, Email: seppo.koskinen@thl.fi, Tel.: +358295248762

60% of the differences. These findings are encouraging since decreasing participation rates are a major problem in population surveys worldwide.

Non-participation; Non-response; Multiple imputation; Inverse probability weighting; Doubly robust methods; Causal model with design.

Classification codes: 62-07; 62-09; 62P10; 62P25

1 Introduction

Unequal sampling probabilities and selective missing data mechanisms markedly complicate the analysis of survey data (14; 35). Due to these challenges, standard tools and analysis methods are not always directly applicable and modifications are required. Making modifications of this kind can easily require tenfold the time and effort, compared to a standard analysis flow in studies with simple random sampling (SRS) and complete data.

In surveys, non-response and non-participation are synonyms, which are used to describe missing data (34). The term "non-response" is commonly used in questionnaire surveys. The term "non-participation" on the other hand is commonly used in epidemiological studies such as health examination surveys (HES), in which study subjects need to arrive to the examination clinic in order to participate in the survey. In this work we use the term non-participation for missing data.

From the population perspective both sampling and non-participation induce missing data: individuals not selected for the sample are missing by design. A decision on sample size is taken during the design phase, on the basis of accuracy targets and budget constraints, and the probabilities associated with the selection process are usually known. By contrast, unit non-response and item non-response lead to unintentional missingness with unknown selection probabilities. These selection probabilities can be estimated only if assumptions such as missing at random (MAR) (46; 41; 51; 37) are feasible or if there is prior knowledge of the selection mechanism (19).

In this paper we aim to present a systematic approach towards the practical handling of missing data. Our motivation is based on our experiences of analysis of the Health 2000 Survey (3) and Health 2011 Survey (31). The Health 2000 Survey was a national health examination survey performed in Finland in 2000-2001. The sample of 9,922 adults was selected by a stratified two-stage cluster sampling design in 2000. In the following we refer to this survey as the baseline. In the Health 2011 Survey, the study subjects covered by the Health 2000 Survey were invited to a health examination. A new sample of young adults was also selected to compensate for the aging of the original cohort. In 2000, the unweighted participation rate was as high as 93% while in 2011 it was only 73%. If the differences in the participation rates are ignored, the studies may not reliably reflect changes in the population's health between 2000 and 2011. The related data collection is described in detail in Section 2.

An analysis of the Health 2000/2011 Survey requires knowledge of complex survey designs (35), longitudinal analysis (40), handling of missing data (41) and model selection methods. To ensure that all aspects of statistical analysis are considered, we divided the analysis flow into the following four steps:

1. Description of data collection using a graphical model.
2. Derivation of the sampling probabilities for the two-stage design.
3. Modeling of non-participation.
4. Systematic comparison of alternative methods of handling missing data.

These steps are explained in Section 3.

On the basis of earlier research on non-participation in health surveys, we know that survey non-participants are more often single men and from younger age groups, with a lower socio-economic status and living in urban areas (8; 9; 13; 22; 30; 39; 54). Non-participants have higher mortality than survey participants (17; 23; 33), more frequent hospitalizations (2; 13; 28), and worse self-reported health (22). They are also known to be daily smokers more frequently (22; 13), to have more mental disorders (9; 39; 29) and to be in receipt of disability benefits more often (39; 30; 29).

Previous studies have reported differences in estimates of population health indicators due to non-participation in the survey results, resulting from the above differences between survey participants and non-participants (61; 38; 55; 16; 15; 11). The non-participation profile in the Health 2000/2011 Survey is presented in Section 4 and discussed in Section 5 of this paper.

By linking the complete survey sample to administrative registers, which have been shown to have a good coverage, we can obtain data on various socio-demographic characteristics as well as on health. We apply these data in two ways. First, the socio-demographic variables are used as auxiliary population information to remove the effects of non-participation.

Second, in order to compare the performance of different methods of handling missing data, we need variables which were observed to the full sample, not only to the participants. By linking the complete survey sample to administrative registers we can obtain health data on disability pension, hospitalization and medicine reimbursements. The prevalence of these variables is estimated using only participants and the various methods. The results are then compared with the prevalences estimated on the basis of the complete survey sample. The results of the comparison are presented in Section 4, and the benefits of the systematic approach are discussed in Section 5.

2 Data and sampling designs

The details of the sample design is described in the subsections below. To summarize the sample designs, the Health 2000 Survey was based on a stratified proportional to size (PPS) two-stage design with health center districts (HCDs)

as primary sampling units (PSUs) and a systematic sample of persons selected at the second stage by the Social Insurance Institution of Finland.

One part of the Health 2011 Survey was simply the full Health 2000 Survey sample, with their given selection probabilities. The 18-28 years olds required an additional sample in order to cover the adult population in 2011.

The inferential populations are (1) the cross-sectional population of those aged 18+ in Finland in 2000, (2) the cross-sectional population of those aged 18+ in 2011, and (3) the cohort of the 2000 population followed to 2011. The following subsections give an overview of the sampling designs, and further details can be found in Supplement A (21).

2.1 The Health 2000 Survey

The Health 2000 Survey was a national health examination survey carried out in 2000–2001. For this survey, the target population encompassed persons aged 18 years or older living in mainland Finland on July 1, 2000.

The design was a stratified two-stage cluster sampling design. 20 geographical strata were based on the 15 largest towns, while the rest of continental Finland was divided into 5 strata based on the university hospital regions.

A total of 80 HCDs were selected for a sample, including the 15 largest towns with probability 1, and a systematic PPS sampling of smaller HCDs (clusters) as the PSUs, in such a way that the sample contained 16 HCDs in each university hospital region. Systematic sampling of persons was applied so that the sample size in each stratum was proportional to the corresponding population base. The total sample size was 9,922. For further details, see Laiho, Djerf and Lehtonen (32).

2.2 The Health 2011 Survey

The sample used for the Health 2011 Survey was designed to provide a representative longitudinal data on the Finnish population. First, the 8,135 eligible sample members from the baseline Health 2000 Survey were invited to participate in the Health 2011 Survey. Of these 8,135 sample members, 1,573 had died, 96 had emigrated and 109 had forbidden any contacts in the future (31) during the 11 year follow-up. In addition, the sample was amended with a new sample of 1,994 young adults aged 18 to 28 years, since the baseline sample had aged by 11 years during the follow-up period. Details of this new sample are presented in the Supplement.

The baseline sample represents the target population on July 1, 2000. It was further realized that during the period up to 2011, the composition of the original sample had changed due to mortality and emigration. Overcoverage detected in the original sample with respect to the 2011 target population was caused by the same mechanism as with respect to the entire population. We therefore have good grounds for considering the composition of the original sample in 2011 to constitute a proper sample of the 2011 target population. However, undercoverage due to immigration to Finland after the year 2000 was

not represented by the baseline sample. The target population therefore became those persons who belonged to the baseline target population and who were alive and resident in Finland in the year 2011. The migration of study subjects between strata was also considered representative with respect to migration within the target population. However, because the allocation of the sample between strata in 2011 could not be controled by means of the sampling design (except in the case of young adults), the stratum sample sizes were considered random.

In addition to the old sample, a new sample of young adults aged 18 to 28 years was sampled. The new sample was selected from the same areas as in the Health 2000 Survey.

2.3 Register data

We had a possibility of linking the full survey sample, to several administrative registers, which have been shown to have a good coverage (see, for example, 62). This was done using the personal ID numbers provided for everyone who resides for at least one year in Finland. The samples were collected by the Population Register Centre using information on the day of birth and municipality of residence as the list variables.

Administrative registers to which the survey data were linked were as follows: 1) Care Register for Health Care (59) (former Hospital Discharge Register). 2) Reimbursement of medical expenses (26) from the Social Insurance Institution of Finland, and 3) Disability benefits and services (25) provided by the Social Insurance Institution of Finland. All these registers are national registers which are regularly updated.

3 Statistical methods

3.1 Description of data collection using a graphical model

The statistical analysis undertaken requires that the study design be expressed in a precise form (Step 1 in the procedure presented in the introduction). For this task, we apply a graphical model extended to present also the study design and the missing data mechanism in a formal way. The concept has been introduced with the name "causal models with design" (24) and it relays on functional causal models defined by Judea Pearl (43; 42). We acknowledge that some readers may have a different view on causality and may not want to talk about causal effects in the absence of factual interventions. These readers may interpret the graphical model as an associational model that gives a useful factorization of the joint distribution. We believe, however, that although our aim is to estimate population statistics without hypotheses of causal relationships, causal considerations are beneficial to understand the study design and the missing data mechanism.

Following Karvanen (24), the nodes of the graphical model are divided into

three classes: causal nodes, selection nodes and data nodes. Here the variables of interest in the population are called causal nodes. These variables are not directly observed but measurements of them are performed in relation to the individuals selected for the sample. The selection nodes function as indicators of sampling and participation, and have the possible values 1 (selected) and 0 (not selected). The selection nodes form a chain where each selection node must have at least one parental selection node. The unique population node r_Ω is an ancestor of all selection nodes and has a value of $r_\Omega = 1$ for all individuals in the population. The data nodes represent actual measurements. For individuals not selected for the sample the measurement is missing. If X is a causal node and r is a selection node, the value of the data node X^* is defined deterministically

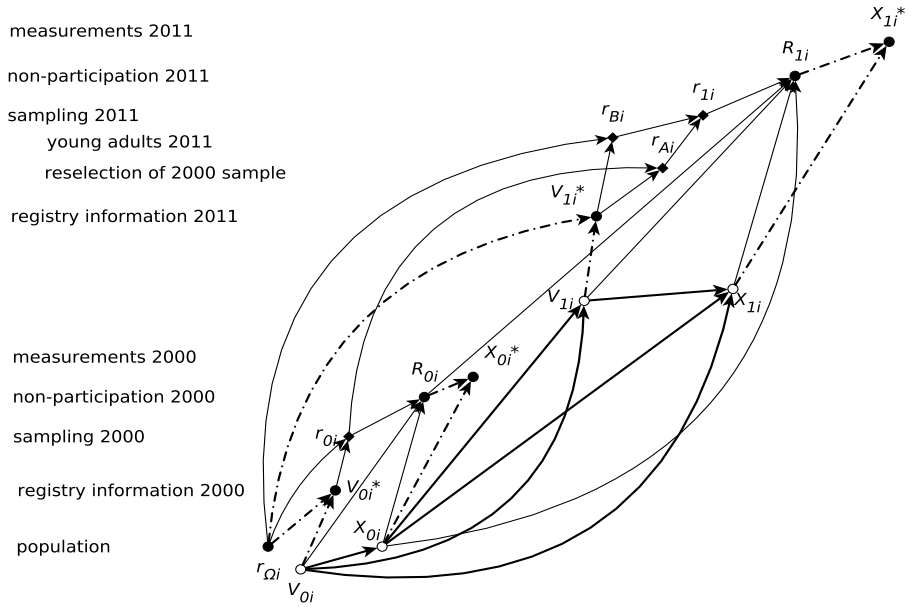
$$X^* := \begin{cases} X, & \text{if } r = 1 \\ \text{NA}, & \text{if } r = 0, \end{cases} \tag{1}$$

where NA stands for missing data.

Figure 1 shows the graphical model for the Health 2000 and 2011 Surveys. The graph seeks to describe the study design and the overall causal/association structure. The causal nodes V_{0i}, X_{0i}, V_{1i} and X_{1i} represent population level variables and the data nodes V^*_{0i}, X^*_{0i}, V^*_{1i} and X^*_{1i} represent the measurements on them. Table 1 presents the notation. The first subscript, 0 or 1, refers to the year, 2000 or 2011, respectively. The second subscript i refers to the individual. Lowercase r denotes a sampling indicator and uppercase R denotes a participation indicator. The selection nodes $r_{\Omega i}$, r_{0i}, r_{Ai}, r_{Bi} and r_{1i} correspond to the sampling design, and R_{0i} and R_{1i} to the missing data mechanism. Index A corresponds to the baseline sample and B to the new sample of young adults in 2011.

The variables (data nodes) of interest can be divided into two groups: strata and registry data available for all individuals in the population, and the covariates measured at the baseline for the participants. Variable V^*_{0i} represents any strata and registry variable in 2000, and X^*_{0i} represents any baseline covariate measured in 2000 for the subject i. Similarly, variable V^*_{1i} represents any strata and registry variable in 2011, and X^*_{1i} represents any baseline covariate measured in 2011 for the subject i. There can be both unit and item non-response in the covariates. The causal/association structures between the registry variables or between the covariates are not specified in the graph because they are not needed in the current analysis.

The population consists of all individuals for whom there was a positive probability of being selected for the study in 2000 or in 2011. This included persons aged 18 years or older and living in mainland Finland on July 1, 2000 and persons aged 18 to 28 years and living in mainland Finland in 2011. A hybrid population of this kind is of technical importance, because all selection probabilities and population distributions are now defined with respect to this population. From the epidemiological perspective we are naturally more interested in the general populations in 2000 and 2011, and changes in health indicators between 2000 and 2011.



- ○ V_{0i} strata and registry data 2000
- ○ X_{0i} covariates 2000
- ● V_{0i}^* observed strata and registry data 2000
- ● X_{0i}^* measured covariates 2000
- ● $r_{\Omega i}$ population indicator
- ◆ r_{0i} sampling indicator 2000
- ● R_{0i} participation indicator 2000
- ○ V_{01} strata and registry data 2011
- ○ X_{1i} covariates 2011
- ● V_{1i}^* observed strata and registry data 2011
- ● X_{1i}^* measured covariates 2011
- ◆ r_{Ai} sampling indicator 2011: reselection of 2000 sample
- ◆ r_{Bi} sampling indicator 2011: young adults
- ◆ r_{1i} sampling indicator 2011
- ● R_{1i} participation indicator 2011

Figure 1: Graphical model with design for Health 2000/2011 Survey. Arrows pointing to selection nodes are solid thin lines, those pointing to data nodes are dash-dotted lines and those pointing to causal nodes are solid, weighted lines. Open circles represent unobserved variables, filled circles represent observed variables and filled diamonds represent decisions made by the researcher. The horizontal axis represents the causal time (the time when the value of a variable in the population is determined) and the vertical axis represents the observational time (the time when the variable becomes available to the researcher).

Table 1: Notation corresponding to the population and sample sizes in stratum s and HCD k, as well as to the observed information and weights for individual i. For the 15 largest towns, define $m^s_\cdot := 1$ and $N^{sk}_{\cdot,[a,b)} := N^s_{\cdot,[a,b)}$. It holds that $S_1 = S_0$ and $m^s_1 = m^s_0$.

	Year 2000	Year 2011
	Age group $[a, b)$	Age group $[a, b)$
Population size	$N_{0,[a,b)}$	$N_{1,[a,b)}$
Sample size	$n^{sk}_{0,[a,b)}$	$n^{sk}_{1,[a,b)}$
Number of strata	S_0	S_1
Population size in s	$N^s_{0,[a,b)}$	$N^s_{1,[a,b)}$
Number of HCD's sampled in s	m^s_0	m^s_1
Population size in s and k	$N^{sk}_{0,[a,b)}$	$N^{sk}_{1,[a,b)}$
Sample size in s and k	$n^{sk}_{0,[a,b)}$	$n^{sk}_{1,[a,b)}$
	All ages	All ages
Selection status	r_{0i}	r_{1i}
Participation status	R_{0i}	R_{1i}
Covariates	X_{0i}	X_{1i}
Register-based covariates	V_{0i}	V_{1i}
Eligibility for invitation		Z_i
Sampling probability	$\mathbb{P}\{\, r_{0i} = 1 \mid V_{0i} \,\}$	$\mathbb{P}\{\, r_{1i} = 1 \mid V_{1i} \,\}$
Participation probability	$\mathbb{P}\{\, R_{0i} = 1 \mid V_{0i},\, X_{0i} \,\}$	$\mathbb{P}\{\, R_{1i} = 1 \mid V_{1i},\, X_{1i} \,\}$
Expansion weight	w_{0i}	w_{1i}

It is assumed that the registry variables V_{0i} precede the baseline covariates. In reality, factors such as health, education and sociodemographic status have a complicated causal structure which can be modelled only if the life courses of the individuals in question are understood in detail. Such modeling is not required for the present analysis.

The probability of being included in the sample $\{i : r_{0i} = 1\}$ depended on the strata variables age and geographical area, which were available from the administrative registries. It is assumed that in 2000 the strata and registry variables V_{0i} and the baseline covariates X_{0i} were associated with the strata and registry variables and the baseline covariates given in the Health 2011 Survey. The register variables included information describing whether the subject was alive and lived in Finland in 2011, and was eligible for invitation to participate in the survey. The sample in 2011 $\{i : r_{1i} = 1\} = \{i : r_{Ai} = 1\} \cup \{i : r_{Bi} = 1\}$ consisted of two groups: the eligible sample members from the Health 2000 Survey $\{i : r_{Ai} = 1\}$ and the new study subjects aged 18 to 28 $\{i : r_{Bi} = 1\}$.

Each subject selected for the sample decided individually whether or not to participate in the survey. We made the MAR assumption that this decision R_{0i} depended on both the observed registry variables and the baseline covariates. The item non-response can be similarly modeled using a separate response indicator for each covariate measurement. In the Health 2011 Survey, each subject selected for the sample also made a decision R_{1i} on whether or not to participate in the survey.

3.2 Sampling probabilities for the two-stage design

In Step 2, the sampling probabilities for 2000 and 2011 are derived from the design. In the terms set in Figure 1, we define the probabilities $\mathbb{P}\{\, r_{0i} = 1 \mid V_{0i}^{*} \,\}$ and $\mathbb{P}\{\, r_{1i} = 1 \mid V_{1i}^{*},\, r_{0i} \,\}$.

The notation required to calculate the sampling probabilities is presented in Table 1. Let us assume that V_{1i} contains V_{0i}. Based on the boundaries in 2000, the true population sizes in both 2011 and 2000 were obtained from Statistics Finland.

In the Health 2000 Survey, the inclusion probability for subject i, who was Age_{0i} years old, belonged to stratum $s := \mathrm{Stratum}_i$ and HCD $k := \mathrm{HCD}_i$, was

defined as

$$\mathbb{P}\{\, r_{0i} = 1 \,|\, \text{Age}_{0i} \,\} = \mathbb{P}\{\, r_{0i} = 1,\, \text{HCD}_i \,|\, \text{Stratum}_i,\, \text{Age}_{0i} \,\} =$$
$$\mathbb{P}\{\, r_{0i} = 1 \,|\, \text{HCD}_i,\, \text{Stratum}_i,\, \text{Age}_{0i} \,\}\, \mathbb{P}\{\, \text{HCD}_i \,|\, \text{Stratum}_i,\, \text{Age}_{0i} \,\} \propto$$

$$\begin{cases} \overbrace{\dfrac{n^{sk}_{0,[30,\infty)}}{N^{sk}_{0,[30,80)} + 2N^{sk}_{0,[80,\infty)}}}^{\text{Probability within HCD } k} \times \dfrac{m^s_0 N^{sk}_{0,[18,\infty)}}{N^s_{0,[18,\infty)}}, & \text{if } \text{Age}_{0i} \in [18, 30) \\ \dfrac{n^{sk}_{0,[30,\infty)}}{N^{sk}_{0,[30,80)} + 2N^{sk}_{0,[80,\infty)}} \times \dfrac{m^s_0 N^{sk}_{0,[18,\infty)}}{N^s_{0,[18,\infty)}}, & \text{if } \text{Age}_{0i} \in [30, 80) \\ \dfrac{2n^{sk}_{0,[30,\infty)}}{N^{sk}_{0,[30,80)} + 2N^{sk}_{0,[80,\infty)}} \times \underbrace{\dfrac{m^s_0 N^{sk}_{0,[18,\infty)}}{N^s_{0,[18,\infty)}}}_{\text{PPS of HCD } k \text{ within stratum } s}, & \text{if } \text{Age}_{0i} \in [80, \infty) \end{cases} \tag{2}$$

It should be noted that if individual i was selected for the sample ($r_{0i} = 1$), then the corresponding HCD was also selected, which explains the first equality in (2). The sampling intervals $(N^{sk}_{0,[30,80)} + 2N^{sk}_{0,[80,\infty)})/n^{sk}_{0,[30,\infty)}$ in the systematic sampling of individuals were equal for the age groups '18 to 29' and '30 to 79' years, and half of those in the age group '80 years and older', making the sampling probabilities twice as high for the oldest age group (12). The sample sizes $n^{sk}_{0,[30,\infty)} = n^{sk}_{0,[30,80)} + n^{sk}_{0,[80,\infty)}$ were equal in each PSU k within stratum s. Within the strata in which element-level sampling was employed the sample sizes were proportional to the corresponding population sizes.

The 'size' of cluster k was the corresponding population size N^{sk}_0 aged 30 or older. The same sampling intervals were used for the '18 to 29 years' age group. The sampling weight was defined as $v_{0i} := 1/\mathbb{P}\{\, r_{0i} = 1 \,|\, \text{Age}_{0i} \,\}$.

All study subjects from the Health 2000 Survey, who were still alive and living in Finland, were invited to participate in the Health 2011 Survey. The sampling probilities for this part of the sample are therefore as follows

$$\mathbb{P}\{\, r_{1i} = 1 \,|\, r_{0i},\, Z_i \,\} = \mathbb{P}\{\, r_{Ai} = 1 \,|\, r_{0i},\, Z_i \,\} = \begin{cases} 1, & \text{if } r_{0i} = 1 \text{ and } Z_i = 1 \\ 0, & \text{otherwise,} \end{cases}$$

where Z_i is a register variable indicating that the study subject is alive and lives in Finland in 2011 ($Z_i = 1$) or has died or left Finland ($Z_i = 0$).

Because the new sample for the '18 to 28 years' age group was created using the same areas as for the Health 2000 Survey, the original sampling probability for cluster k was the same in 2011 as in 2000. The PPS probabilities in 2011 were the same as in 2000, due to which the inclusion probabilities for this part

of the sample in 2011 were

$$\mathbb{P}\{\, r_{1i} = 1 \,|\, \mathrm{Age}_{1i},\, V_{1i} \,\} = \mathbb{P}\{\, r_{1i} = 1,\, \mathrm{HCD}_i \,|\, \mathrm{Stratum}_i,\, \mathrm{Age}_{1i},\, V_{1i} \,\} =$$
$$\mathbb{P}\{\, r_{1i} = 1 \,|\, \mathrm{HCD}_i,\, \mathrm{Stratum}_i,\, \mathrm{Age}_{1i},\, V_{1i} \,\}\, \mathbb{P}\{\, \mathrm{HCD}_i \,|\, \mathrm{Stratum}_i,\, \mathrm{Age}_{0i} \,\} \propto$$
$$\overbrace{\frac{n^{sk}_{1,[18,29)}}{N^{sk}_{1,[18,29)}}}^{\text{Probability within HCD } k \text{ in year 2011}} \times \underbrace{\frac{m^s_0 N^{sk}_{0,[18,\infty)}}{N^s_{0,[18,\infty)}}}_{\text{PPS of HCD } k \text{ within stratum } s \text{ in year 2000}}$$
$$\text{if } \mathrm{HCD}_i = k,\, \mathrm{Stratum}_i = s,\, \mathrm{Age}_{1i} \in [18, 29) \quad (3)$$

3.3 Modeling non-participation

In Step 3, the conditional participation probabilities for 2000 and 2011 are estimated based on the related data. Based on Figure 1, here we estimate the probabilities $\mathbb{P}\{\, R_{0i} = 1 \,|\, r_{0i} = 1,\, V_{0i},\, X_{0i} \,\}$ and $\mathbb{P}\{\, R_{1i} = 1 \,|\, r_{1i} = 1,\, R_{0i},\, V_{1i},\, X_{1i} \,\}$.

Where the participation rates are as high as in the Health 2000 Survey, it would be common practice to use the calibration weights of the Health 2000 Survey (12). We refer to these as the baseline weights below. For the year 2000, the baseline weights are derived from equation (2), with no further adjustments for non-participation. It is assumed that weighted statistics such as means and prevalences provide representative results on the target population. In addition to the cross-sectional statistics, follow-up data can also be analyzed using baseline weights and the methods applied for cohort studies.

Formally, the participation probability can be stated as

$$\mathbb{P}\{\, R_{0i} = 1 \,|\, V_{0i},\, X_{0i} \,\} =$$
$$\mathbb{P}\{\, R_{0i} = 1 \,|\, r_{0i} = 1,\, V_{0i},\, X_{0i} \,\}\, \mathbb{P}\{\, r_{0i} = 1 \,|\, V_{0i} \,\} \propto \mathbb{P}\{\, R_{0i} = 1 \,|\, V_{0i} \,\}\,. \quad (4)$$

The proportionality in (4) holds in the case that the MAR assumption for the non-participation is valid. Non-response was accounted for using the calibration weights based on language, gender, age and area (12), thus

$$\mathbb{P}\{\, R_{0i} = 1 \,|\, V_{0i},\, X_{0i} \,\} \approx \mathbb{P}\{\, R_{0i} = 1 \,|\, V^*_{0i} \,\}\,.$$

In 2011, participation rates were significantly lower and the means and prevalences of the population in 2011 cannot be reliably estimated using the baseline weights. From Figure 1 we can see that it is also assumed that the decision on participating R_{1i} in the Health 2011 Survey depends on the health status and other covariates $(V_{1i},\, X_{1i})$ which the survey is intended to measure. This is a missing-not-at-random (MNAR) situation; in general, the participation probabilities cannot be estimated on the basis of the data only. However, the baseline covariates measured in 2000 include information on various risk and lifestyle factors predicting future diseases or functional disabilities, which are expected to be common causes of non-participation in 2011.

We divide the sample in 2011 into two parts. The first part consists of the participants of the Health 2000 Survey ($R_{0i} = 1$). We assume that the participation probabilities in 2011 can be modeled reasonably well using the registry variables in 2011 and the covariates measured in 2000

$$\begin{aligned}\mathbb{P}\{\, R_{1i} = 1 \mid r_{Ai} = 1,\, V_{1i},\, R_{0i} = 1,\, X_{0i},\, X_{1i}\,\} \\ \approx \mathbb{P}\{\, R_{1i} = 1 \mid r_{Ai} = 1,\, V^*_{1i},\, R_{0i} = 1,\, X^*_{0i}\,\}. \quad (5)\end{aligned}$$

We model the latter probability using the logistic regression model

$$\operatorname{logit}\left(\mathbb{P}\{\, R_{1i} = 1 \mid r_{Ai} = 1, V^*_{1i},\, R_{0i} = 1,\, X^*_{0i}\,\}\right) = \alpha_1 V^*_{1i} + \beta X^*_{0i}, \quad (6)$$

where α_1 and β are regression parameters.

The second part consists of the small group of subjects who did not participate in the Health 2000 Survey ($R_{0i} = 0$) and the new sample of young adults. They are handled similarly, but using only the register-based information V_{1i}.

$$\operatorname{logit}(\mathbb{P}\{\, R_{1i} = 1 \mid r_{Ai} = 1, V^*_{1i},\, r_{1i} = 1,\, R_{0i} = 0\,\}) = \alpha_0 V^*_{1i} + \alpha_R \quad (7)$$

$$\operatorname{logit}(\mathbb{P}\{\, R_{1i} = 1 \mid r_{Bi} = 1, V^*_{1i},\, r_{0i} = 0\,\}) = \alpha_0 V^*_{1i}, \quad (8)$$

where α_R describes the effect of non-participation in 2000 on participation in 2011. The regression models corresponding to the observed register data V^*_{1i} (and α_0 and α_1) depicted the interactions between age group, gender and education.

Combining the above results, the probability of participation for each individual i can now be expressed as

$$\begin{aligned}&\mathbb{P}\{\, R_{1i} = 1 \mid V_{1i},\, r_{0i},\, R_{0i},\, X_{0i}\,\} \propto \\ &\qquad \mathbb{P}\{\, R_{1i} = 1 \mid r_{1i} = 1,\, V^*_{1i},\, R_{0i},\, X^*_{0i}\,\}\,\mathbb{P}\{\, r_{1i} = 1 \mid V^*_{1i}\,\} = \\ &\begin{cases} \mathbb{P}\{\, R_{1i} = 1 \mid r_{1i} = 1,\, V^*_{1i},\, R_{0i},\, X^*_{0i}\,\}, & \text{if } r_{0i} = 1 \text{ and } Z_i = 1 \\ \mathbb{P}\{\, R_{1i} = 1 \mid r_{1i} = 1,\, V^*_{1i},\, r_{0i} = 0,\, X^*_{0i}\,\} \times & \text{if } r_{0i} = 0 \text{ and } \mathrm{Age}_{1i} \in [18, 29) \\ \qquad \mathbb{P}\{\, r_{Bi} = 1 \mid V^*_{1i}\,\}, & \end{cases} \\ &= \begin{cases} \operatorname{logit}^{-1}(\alpha_1 V^*_{1i} + \beta X^*_{0i}), & \text{if } r_{0i} = 1, R_{0i} = 1 \text{ and } Z_i = 1, \\ \operatorname{logit}^{-1}(\alpha_0 V^*_{1i} + \alpha_R), & \text{if } r_{0i} = 1, R_{0i} = 0 \text{ and } Z_i = 1, \\ \operatorname{logit}^{-1}(\alpha_0 V^*_{1i})\mathbb{P}\{\, r_{Bi} = 1 \mid V_{1i}\,\}, & \text{if } r_{0i} = 0 \text{ and } \mathrm{Age}_{1i} \in [18, 29), \\ 0, & \text{otherwise} \end{cases} \quad (9)\end{aligned}$$

where the sampling probability $\mathbb{P}\{\, r_{Bi} = 1 \mid V^*_{1i}\,\}$ is given in equation (3).

For this work we selected numerous variables covering various aspects of the information collected in the Health 2000 Survey. Together with age and gender, and with or without the interactions between the variable and age and/or gender, these variables were then entered one at a time into logistic regression models. The Bayesian information criterion (BIC, 50) was then applied to assess which variables had better predictive power with respect to non-participation

than the model accounting only for age group and gender. The best predictors were then entered as main effects into the same model. Using the Wald test, the variables, whose p-value was below 0.20 were selected for the final model. This approach was compared with the BIC applied in the multivariate models, which contained the predictors selected by the univariate BIC described above. Due to the item non-response, these steps were conducted using the complete-case data ignoring the sampling design. See Section 4 for a description of the selected model.

Diagnostics of the weighting model using the le Cessie – van Houwelingen – Copas – Hosmer unweighted sum of squares test for global goodness of fit (20) gave p-value 0.85. A large p-value indicates that the model cannot be disqualified with the data.

3.4 Comparison of missing data methods

The models derived above are used for Step 4. Commonly applied methods in handling missing data are the inverse probability weighted (IPW) analysis (7; 45), doubly robust (DR) methods (5) and multiple imputation (MI) (48; 49). The MI methods have previously been applied, for example, in longitudinal data analyses (56) and survey analyses (18). More specifically, we use the MI method called the multivariate imputation using chained equations (64; 63). All these methods apply the sampling weights described in subsection 3.2 to provide results which represent the population. The IPW and DR methods also require weights which handle the non-participation and which were defined in subsection 3.3. As these methods are widely known and they are applied here using generally available software, we describe the details in the Supplement B.1 (21). Three imputation models using the predictive mean matching (PMM, 47; 36) method are applied, and the imputed data sets are analyzed using the baseline weights (53). MI1 contains only the register-based sociodemographic variables, MI2 the same variables as IPW and DR methods above, and MI3 in addition to those variables contained in MI2, also the biological risk factors measured at the baseline survey. In addition to the PMM method, also the recursive partitioning and regression trees (RPART, (58), option `cart` in package `mice`) were applied to two other MI models. MI4 contained the same variables as MI3. MI5 contained all variables selected by the Akaike information criterion (AIC, 1) applied to age and gender adjusted univariate logistic regression models using complete-case data without sampling design or weights. As a result, MI5 contained a much larger group of variables than the other imputation models. Supplement Table I presents the selected additional variables. Neither the sampling design nor the weights were not applied in the MI procedures. 50 imputed data sets were generated as it has been suggested that the number of imputations should be greater than the percentage of missing data (66).

Although alternative missing data methods were available, in this work we concentrate on these methods. Bayesian inference has benefits in terms of handling missing data, since the uncertainties involved in both the parameters and data can be properly accounted for in the predictive distributions of the missing

Table 2: Sample sizes, and participation rates in any part of the survey or in the health examination (HE) part in different age groups of the Health 2011 Survey.

	..... Age groups				
	18-28	29-49	50-74	75+	All
Total sample	1,994	3,306	3,840	989	10,129
Participation (%)	42	68	79	65	67
HE sample	415	3,306	3,840	989	8,550
HE participation (%)	29	50	66	54	57

data values. Tanner and Wong (57) introduced the data augmentation method, in which the missing data are integrated out from the joint distribution of the data and the model parameters. The posterior distribution of the parameters can then be obtained from the resulting distribution. In practice, this integration can be handled numerically using the Markov chain Monte Carlo (MCMC) methods (44) and, for example, the OpenBugs software (60).

Various methods of handling missing data were applied to calculating prevalences adjusted for non-participation. The true prevalences of the disability benefits in 2009, hospitalizations in 2010 and reimbursement of medications in 2011, which were available from the administrative registeries for all sample members, were calculated as the sampling weighted prevalences using the full sample. When assessing the goodness of the missing data methods, the values of these outcome variables were set missing to the non-participants before imputation or calibration of weights.

Younger individuals under the age of 30 years were excluded since these health-related events are rare, entailing that the prevalences are likely to be close to zero and making differences between various methods difficult to detect.

4 Results

Participation rates were much lower in the youngest age group than in the '50 to 74 years' age group. They were also lower in the oldest age group (Table 2). The participation rate for the health examination (HE) was lower than the participation rate in any part of the survey including the questionnaires, interviews and HE (total). 8,135 sample members of the Health 2000 Survey were eligible to the Health 2011 Survey. Of the 5,903 participants in 2011, 5,602 (94.9%) had participated in 2000. Of the 2,232 nonparticipants in 2011, 1,589 (71.2%) had participated in 2000.

In 2000, participants appeared to have higher hospitalization prevalences than non-participants under 65 years, but prevalences were lower among older participants (Supplement Table II, 21). In 2011 these differences appeared only in those older than 65. Non-participants had higher disability pension prevalences in 2011.

Univariate model selection demonstrated that the self-reported work ability

score (together with the interaction between age group, gender and education) was the best predictor for the response in 2011: BIC=5331. The null model containing only the interaction between age group, gender and education gave the following result: BIC=5337. A better BIC value than the null model was also recorded for the following predictors: language (Finnish vs. Swedish or other); participation frequency in clubs or associations; self-reported health status; and the work ability. When one or more of these predictors were entered into the regression model simultaneously, both the model selection and the Wald tests showed that the best results were achieved for the work ability, language and participation frequency in clubs or associations in addition to the interaction of age group, gender and education (BIC=5325).

The participation rate was very low among men with low educational attainment and amongst the '29 to 34 years' age group as shown by the OR estimates of the main effects (Supplement Table III, 21). Poor self-reported work ability, language other than Finnish, and either no or high activity in clubs or associations also decreased participation in 2011.

Accompanied by the baseline weights, multiple imputation based on the chained equations appeared to remove almost all differences between full sample estimates and estimates based on the participants prevalences (Table 3). The soundness of the imputation methods and models varied depending on the outcome. The RPART method was the best in two out of three outcomes and PMM in one. The large imputation model (MI5) was best only for one outcome. In two outcomes out of three, the estimate based on the large imputation model (MI3) was the closest to the true prevalence. The estimated standard errors were the smaller the larger the imputation model was. These results are in accordance with the general fact that the more information is available for predicting the missing values, the more accurate the estimates and the smaller the standard errors are. However, there seemed to be almost no benefits in adding a large number of additional variables based on the AIC in MI5. Baseline weights had almost no influence at the baseline, nor did they remove the differences for the 2011 results. Adjusted for the non-participation in 2011, the weights improved the estimates considerably by removing more than half of the differences. The doubly robust method appeared to provide even better results by removing at least 60% of the differences for all three outcome variables. In the Health 2000 Survey, the participation rate was very high, thus the differences were small.

Table 3: Comparison of different methods in order to correct for the effects of missing data in the age group '30 years and older'. True prevalences are those given in the sample. With respect to the missing data methods, baseline weights correspond to the Health 2000 Survey calibration weights, resurvey weights to the Health 2011 Survey weights and HE weights to the Health 2011 Survey weights based on participation in the health examination (HE). Multiple imputation results MI1, MI2 and MI3 based on PMM refer to the imputation models containing the register-based sociodemographic variables, as well as self-reported work ability, health and participation frequency in clubs or associations, and biological risk factors. MI4 and MI5 based on RPART refer to imputation models MI3 and a large imputation model selected using the AIC, respectively. Clustering of the sampling design was accounted for in the analysis ("complex") or not ("SRS").

Variable	Clustering	Missing data method	Year	Prev. (%)	SE
Disability pension	SRS	None	2000	6.59	0.32
	Complex	Baseline weights	2000	6.69	0.37
	Complex	True prevalence	2000	6.61	0.31
	SRS	None	2011	8.74	0.41
	Complex	Baseline weights	2011	8.92	0.42
	Complex	Resurvey weights	2011	9.21	0.44
	Complex	Resurvey HES weights	2011	8.47	0.50
	Complex	Doubly Robust	2011	9.24	0.49
	Complex	Baseline weights; MI1	2011	9.26	0.51
	Complex	Baseline weights; MI2	2011	9.59	0.47
	Complex	Baseline weights; MI3	2011	9.47	0.44
	Complex	Baseline weights; MI4	2011	9.33	0.44
	Complex	Baseline weights; MI5	2011	9.31	0.44
	Complex	True prevalence	2011	9.53	0.37
Hospitalization	SRS	None	2000	14.46	0.45
	Complex	Baseline weights	2000	14.42	0.44
	Complex	True prevalence	2000	14.02	0.44
	SRS	None	2011	16.63	0.54
	Complex	Baseline weights	2011	16.57	0.52
	Complex	Resurvey weights	2011	16.92	0.54
	Complex	Resurvey HES weights	2011	16.59	0.65
	Complex	Doubly Robust	2011	17.16	0.58
	Complex	Baseline weights; MI1	2011	17.26	0.63
	Complex	Baseline weights; MI2	2011	17.12	0.57
	Complex	Baseline weights; MI3	2011	17.25	0.55
	Complex	Baseline weights; MI4	2011	17.32	0.54
	Complex	Baseline weights; MI5	2011	17.53	0.55
	Complex	True prevalence	2011	17.63	0.48
Reimbursement	SRS	None	2000	25.56	0.56
	Complex	Baseline weights	2000	25.13	0.64
	Complex	True prevalence	2000	25.14	0.54
	SRS	None	2011	40.14	0.71
	Complex	Baseline weights	2011	40.34	0.78
	Complex	Resurvey weights	2011	40.79	0.79

Complex	Resurvey HES weights	2011	40.37	0.87
Complex	Doubly Robust	2011	41.27	0.74
Complex	Baseline weights; MI1	2011	40.81	0.91
Complex	Baseline weights; MI2	2011	41.63	0.81
Complex	Baseline weights; MI3	2011	41.64	0.75
Complex	Baseline weights; MI4	2011	41.72	0.73
Complex	Baseline weights; MI5	2011	41.59	0.76
Complex	True prevalence	2011	41.88	0.62

5 Discussion

The systematic approach we have presented, for the analysis of complex study design with missing data, has several benefits over less systematic, ad hoc approaches. Causal models with design force statisticians to make the study design and the missing data mechanism explicit, although in many population surveys hypotheses concerning causal relationships are not relevant. In such association studies these models can be considered as graphical models describing both associations between variables and the temporal ordering of variables. This helps statisticians to think logically and makes it easier for the reader to follow the presentation. The need to communicate the study design accurately is also stressed in the reporting guidelines such as the STROBE Statement (65). Using the systematic approach, the complex likelihood can be split into parts of manageable size. Most importantly, sampling probabilities and participation probabilities are considered separately, since the former are known by their design and the latter are estimated from the data. This helps us to see the common factors between the alternative missing data methods and to observe the differences between these methods.

The systematic approach was beneficial to integrating the complex sampling design and non-participation within a single probabilistic framework during the analysis of the Health 2000/2011 data. This was particularly useful in estimating the weights for the IPW method. NMAR based on possibly unobservable causal nodes could be a hypothetical model for non-participation, whereas the actual estimable model for non-participation assuming MAR is based on observations (data nodes). The systematic approach reveals where assumptions have been made in order to render the model identifiable. These assumptions can also be communicated to other researchers using graphs as in Figure 1.

Gelman (14) compared weighting and modeling as possible solutions to handling survey design and non-participation in statistical analyses. While both approaches had advantages and disadvantages, in our case the potential complexity involved in constructing the weights was at least partially realized. In a multi-wave survey such as the Health 2011 Survey, in which each wave involved several parts such as the health examination, interviews and questionnaires, the definition of participation alone becomes complicated. A different set of weights would be needed for each definition of this kind. Multiple imputation can handle not only different participation profiles, but also item non-participation.

The imputed data can then be analyzed using survey analysis methods to handle different sampling probabilities, unit non-response and clustering as we have done. This approach can be a partial solution to achieving the aim expressed by Gelman (14): "Our ideal procedure should be as easy to use as hierarchical modeling, with population information included using poststratification." Related to hierarchical modeling, we did not account for intra-cluster correlations, because we anticipated that migration during the 11-year follow-up period would have eroded such associations, but this can create bias in variance estimates (27).

In the original sample for 2000, 109 persons (around 1% of the sample) specifically forbade any further contacts after the Health 2000 Survey. However, due to the written consent in 2000, register-based follow-up and utilization of the baseline data are possible thus allowing the missing data analyses.

As we have demonstrated in the case of health-related outcomes, which are often associated with non-participation, various statistical methods to handle missing data appear to be effective in reducing the differences between full sample estimates and estimates based on the participants. In particular, multiple imputation with chained equations appeared to be the best method in our comparison. In addition, the doubly robust method provided good results. These findings are encouraging, since decreasing participation rates have been a major problem in population surveys worldwide. For practical work, tools for statistical analyses based on multiple imputation are available in many statistical software packages, for example in SAS, Stata, SPSS and R, whereas tools intended for the application of doubly robust methods are not commonly available.

Our results demonstrate that the more baseline information can be utilized in the MI, the more accurate results can be obtained. In our case the participation rates were exceptionally high at the baseline, but it is likely that similar benefits can be obtained also in other multi-wave surveys with decreasing participations rates. Furthermore, in multi-wave surveys information on individuals, who did not participate in the early waves but participated in the later ones, can also be utilized in the MI. Multi-wave surveys can provide more information, which improves the accuracy of the results compared to those obtained by only cross-sectional surveys.

It has been shown that a weighting model should contain good predictors of the outcome rather than the missingness (4; 52; 6). Our aim in the Health 2011 Survey has been to provide researchers general-purpose tools (based on the IPW as weights are easy to use also for non-statisticians) to handle missing data. As there are a couple of thousand variables aimed to cover a large variety of lifestyle and risk factors as well as health and other outcomes, we cannot provide optimized weights for all research questions, thus we concentrated on variables with predictive power on the missingness. We compared five imputation models based on different amount of register or baseline survey information, which improved the results. The imputation model MI2 utilized basically the same information as the resurvey weights, but the MI2 results were slightly closer to the true prevalences. The MI3 contained also the BMI, systolic blood pressure and smoking (measured in 2000), which are important risk factors to various health outcomes such as the ones we have applied in our work. It was also

notable that the MI3 model, which contained the important predictors of health outcomes, has slightly smaller standard errors, which in accordance with earlier results (52; 10) that inclusion of auxiliary variables associated with the outcome in the imputation model can improve accuracy of the estimates. The differences between MI3, which appeared to perform best, and other methods were not very large, however. Results based on the RPART method (MI4 and MI5) were practically as good as MI2 and MI3. MI5 contained a considerably larger number of variables selected using the AIC than the other imputation or weighting models, but it did not perform better than the other imputation models, thus a relatively small number of variables in the imputation model seemed to be sufficient in our case.

Many similarities with the results reported in the literature can be observed in the patterns of non-participation in the Health 2011 Survey. In the Health 2011 Survey, the participation rates were lower among men, younger age groups and those with lower educational attainments. Similar results have been reported for other studies (8; 9; 13; 22; 30; 39; 54). In addition, our observation that non-participants are more often in receipt of disability pensions than participants echoes previous reports from Finland (30), Norway (29) and Sweden (39).

Acknowledgements

This work was supported by the Academy of Finland under Grant 266251.
The authors have no relevant financial or nonfinancial relationships to disclose.
The Version of Record of this manuscript has been published and is available in Journal of Applied Statistics, February 20 2016
http://www.tandfonline.com/10.1080/02664763.2016.1144725.

Supplemental material

Details of the sampling designs and missing data methods. Two additional tables.

References

[1] H. Akaike, *Maximum likelihood identification of gaussian autoregressive moving average models*, Biometrika 60 (1973), pp. 255–265.

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

[3] A. Aromaa and S. Koskinen (eds.), *Health and Functional Capacity in Finland: Baseline Results of the Health 2000 Health Examination Survey*,

no. 12 in B, National Public Health Institute, 2004, Available at http://terveys2000.fi/julkaisut/baseline.pdf.

[4] P.C. Austin, *The performance of different propensity-score methods for estimating relative risks*, Journal of clinical epidemiology 61 (2008), pp. 537–545.

[5] H. Bang and J.M. Robins, *Doubly robust estimation in missing data and causal inference models*, Biometrics 61 (2005), pp. pp. 962–972, Available at http://www.jstor.org/stable/3695907.

[6] M.A. Brookhart, S. Schneeweiss, K.J. Rothman, R.J. Glynn, J. Avorn, and T. Stürmer, *Variable selection for propensity score models*, American Journal of Epidemiology 163 (2006), pp. 1149–1156.

[7] C.M. Cassel, C.E. Sarndal, and J.H. Wretman, *Some uses of statistical models in connection with the nonresponse problem*, in *Incomplete data in sample surveys*, W.G. Madow and I. Olkin, eds., Vol. 3, Academic Press, New York, 1983, pp. 143–160.

[8] P. Chou, H.S. Kuo, C.H. Chen, and H.C. Lin, *Characteristics of non-participants and reasons for non-participation in a population survey in kin-hu, kinmen*, European Journal of Epidemiology 13 (1997), pp. 195–200.

[9] G. Cohen and J.C. Duffy, *Are nonrespondents to health surveys less healthy than respondents?*, Journal of Official Statistics 18 (2002), pp. 13–23.

[10] L.M. Collins, J.L. Schafer, and C.M. Kam, *A comparison of inclusive and restrictive strategies in modern missing data procedures.*, Psychological methods 6 (2001), p. 330.

[11] M.H. Criqui, M. Austin, and E. Barrett-Connor, *The effect of non-response on risk ratios in a cardiovascular disease study*, Journal of Chronic Diseases 32 (1979), pp. 633–638.

[12] K. Djerf, J. Laiho, R. Lehtonen, T. Härkänen, and P. Knekt, *Methodology report: Health 2000 Survey*, chap. Weighting and statistical analysis, no. 26 in B, KTL (2008), pp. 182–200, Available at http://www.terveys2000.fi/doc/methodologyrep.pdf.

[13] T. Drivsholm, L.F. Eplov, M. Davidsen, T. Jörgensen, H. Ibsen, H. Hollnagel, and K. Borch-Johnsen, *Representativeness in population-based studies: a detailed description of non-response in a Danish cohort study*, Scandinavian Journal of Public Health 34 (2006), pp. 623–631.

[14] A. Gelman, *Struggles with survey weighting and regression modeling*, Statistical Science 22 (2007), pp. 153–164.

[15] K.M. Gorey and M. Trevisan, *Secular trends in the United States black/white hypertension prevalence ratio: Potential impact of diminishing response rates*, American Journal of Epidemiology 147 (1998), pp. 95–99.

[16] J. Gundgaard, O. Ekholm, E.H. Hansen, and N.K. Rasmussen, *The effect of non-response on estimates of health care utilisation: linking health surveys and registers*, The European Journal of Public Health 18 (2008), pp. 189–194.

[17] M. Hara, S. Sasaki, T. Sobue, S. Yamamoto, and S. Tsugane, *Comparison of cause-specific mortality between respondents and nonrespondents in a population-based prospective study: ten-year follow-up of JPHC study cohort i*, Journal of Clinical Epidemiology 55 (2002), pp. 150–156.

[18] Y. He, A.M. Zaslavsky, M.B. Landrum, D.P. Harrington, and P. Catalano, *Multiple imputation in a large-scale complex survey: a practical guide.*, Stat Methods Med Res 19 (2010), pp. 653–670, Available at `http://dx.doi.org/10.1177/0962280208101273`.

[19] J.J. Heckman, *Sample selection bias as a specification error*, Econometrica: Journal of the Econometric Society (1979), pp. 153–161.

[20] D.W. Hosmer, T. Hosmer, S. Le Cessie, S. Lemeshow, *et al.*, *A comparison of goodness-of-fit tests for the logistic regression model*, Statistics in medicine 16 (1997), pp. 965–980.

[21] T. Härkänen, J. Karvanen, H. Tolonen, R. Lehtonen, K. Djerf, T. Juntunen, and S. Koskinen, *Supplement to "systematic handling of missing data in complex study designs - experiences from the health 2000 and 2011 surveys"*. .

[22] R. Jackson, L.E. Chambless, K. Yang, T. Byrne, R. Watson, A. Folsom, E. Shahar, and W. Kalsbeek, *Differences between respondents and nonrespondents in a multicenter community-based study vary by gender and ethnicity*, Journal of Clinical Epidemiology 49 (1996), pp. 1441–1446.

[23] P. Jousilahti, V. Salomaa, K. Kuulasmaa, M. Niemela, and E. Vartiainen, *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 59 (2005), pp. 310–315.

[24] J. Karvanen, *Study design in causal models*, Scandinavian Journal of Statistics 42 (2015), pp. 361–377, Available at `http://arxiv.org/abs/1211.2958`.

[25] Kela, *Statistics on disability benefits and services provided by kela* (2015), Available at `http://www.kela.fi/web/en/statistics-by-topic_statistics-on-disability-benefits-provided-by-kela`.

[26] Kela, *Statistics on reimbursement entitlements in respect of medicines* (2015), Available at `http://www.kela.fi/web/en/statistics-by-topic_statistics-on-reimbursement-entitlements-in-respect-of-medicines`.

[27] J.K. Kim, J. Michael Brick, W.A. Fuller, and G. Kalton, *On the bias of the multiple-imputation variance estimator in survey sampling*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (2006), pp. 509–521.

[28] M. Kjoller and H. Thoning, *Characteristics of non-response in the Danish health interview surveys, 1987-1994*, The European Journal of Public Health 15 (2005), pp. 528–535.

[29] A.K. Knudsen, M. Hotopf, J.C. Skogen, S. Averland, and A. Mykletun, *The health status of nonparticipants in a population-based health study the Hordaland health study*, American Journal of Epidemiology 172 (2010), pp. 1306–1314.

[30] K. Korkeila, S. Suominen, J. Ahvenainen, A. Ojanlatva, P. Rautava, H. Helenius, and M. Koskenvuo, *Non-response and related factors in a nationwide health survey*, European Journal of Epidemiology 17 (2001), pp. 991–999.

[31] S. Koskinen, A. Lundqvist, and N. Ristiluoma (eds.), *Terveys, toimintakyky ja hyvinvointi Suomessa 2011*, no. 68 in Raportti, National Institute for Health and Welfare, 2012, Available at `http://urn.fi/URN:ISBN:978-952-245-769-1`.

[32] J. Laiho, K. Djerf, and R. Lehtonen, *Methodology report: Health 2000 Survey*, chap. Sampling design, no. 26 in B, National Public Health Institute (2008), pp. 13–15, Available at `http://www.terveys2000.fi/doc/methodologyrep.pdf`.

[33] S.B. Larsen, S.O. Dalton, J. Schüz, J. Christensen, K. Overvad, A. Tönneland, C. Johansen, and A. Olsen, *Mortality among participants and nonparticipants in a prospective cohort study*, European Journal of Epidemiology 27 (2012), pp. 837–845.

[34] J.M. Last, *A Dictionary of Epidemiology*, 4th ed., Oxford University Press, New York, 2001.

[35] R. Lehtonen and E. Pahkinen, *Practical methods for design and analysis of complex surveys*, Wiley. com, 2004.

[36] R.J. Little, *Missing-data adjustments in large surveys*, Journal of Business & Economic Statistics 6 (1988), pp. 287–296.

[37] R.J. Little and N. Zhang, *Subsample ignorable likelihood for regression analysis with missing data*, Journal of the Royal Statistical Society: Series C (Applied Statistics) 60 (2011), pp. 591–605.

[38] A.J.V. Loon, M. Tijhuis, H.S. Picavet, P.G. Surtees, and J. Ormel, *Survey non-response in the netherlands: effects on prevalence estimates and associations*, Annals of Epidemiology 13 (2003), pp. 105–110.

[39] I. Lundberg, K.D. Thakker, T. Hallstrom, and Y. Forsell, *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 40 (2005), pp. 475–483.

[40] P. Lynn, *Methodology of longitudinal surveys*, John Wiley & Sons, 2009.

[41] G. Molenberghs and M.G. Kenward, *Missing Data in Medical Research*, Wiley, 2007.

[42] J. Pearl, *Causal inference in statistics: An overview*, Statistics Surveys 3 (2009), pp. 96–146, Available at `http://projecteuclid.org/euclid.ssu/1255440554`.

[43] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009.

[44] C.P. Robert and G. Casella, *Monte Carlo statistical methods*, 2nd ed., Springer Verlag New York, 2004.

[45] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Estimation of regression coefficients when some regressors are not always observed*, Journal of the American Statistical Association 89 (1994), pp. 846–866, Available at `http://amstat.tandfonline.com/doi/full/10.1080/01621459.1994.10476818`.

[46] D.B. Rubin, *Inference and missing data*, Biometrika 63 (1976), pp. 581–592.

[47] D.B. Rubin, *Statistical matching using file concatenation with adjusted weights and multiple imputations*, Journal of Business & Economic Statistics 4 (1986), pp. 87–94.

[48] D.B. Rubin, *Multiple imputation for nonresponse in surveys*, Wiley, New York, 1987.

[49] J.L. Schafer, *Multiple imputation: a primer.*, Statistical Methods in Medical Research 8 (1999), pp. 3–15.

[50] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics 6 (1978), pp. pp. 461–464, Available at `http://www.jstor.org/stable/2958889`.

[51] S. Seaman, J. Galati, D. Jackson, and J. Carlin, *What is meant by "missing at random"*, Statistical Science 28 (2013), pp. 257–268.

[52] S.R. Seaman and I.R. White, *Review of inverse probability weighting for dealing with missing data*, Statistical Methods in Medical Research 22 (2013), pp. 278–295.

[53] S.R. Seaman, I.R. White, A.J. Copas, and L. Li, *Combining multiple imputation and inverse-probability weighting*, Biometrics 68 (2012), pp. 129–137.

[54] A.J. Sögaard, R. Selmer, E. Bjertness, and D. Thelle, *The Oslo health study: The impact of self-selection in a large, population-based survey*, International Journal for Equity in Health 3 (2004), p. 3.

[55] E. Shahar, A.R. Folsom, and R. Jackson, *The effect of nonresponse on prevalence estimates for a referent population: insights from a population-based cohort study*, Annals of Epidemiology 6 (1996), pp. 498–506.

[56] M. Spratt, J. Carpenter, J.A. Sterne, J.B. Carlin, J. Heron, J. Henderson, and K. Tilling, *Strategies for multiple imputation in longitudinal studies*, American journal of epidemiology 172 (2010), pp. 478–487.

[57] M.A. Tanner and W.H. Wong, *The calculation of posterior distributions by data augmentation*, The Journal of the American Statistical Association 82 (1987), pp. 528–550.

[58] T. Therneau, B. Atkinson, and B. Ripley, *rpart: Recursive Partitioning and Regression Trees* (2015), Available at `http://CRAN.R-project.org/package=rpart`, r package version 4.1-10.

[59] THL, *Care register for health care* (2015), Available at `http://www.thl.fi/en_US/web/en/statistics/information/register_descriptions/careregister_healthcare`.

[60] A. Thomas, B. O'Hara, U. Ligges, and S. Sturtz, *Making BUGS open*, R News 6 (2006), pp. 12–17.

[61] H. Tolonen, A. Dobson, and S. Kulathinal, *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 20 (2005), pp. 887–898.

[62] A.M. Tolppanen, H. Taipale, M. Koponen, P. Lavikainen, A. Tanskanen, J. Tiihonen, and S. Hartikainen, *Use of existing data sources in clinical epidemiology: Finnish health care registers in Alzheimer's disease research–the medication use among persons with Alzheimer's disease (MEDALZ-2005) study*, Clinical Epidemiology 5 (2013), p. 277.

[63] S. van Buuren and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in R*, Journal of Statistical Software 45 (2011), pp. 1–67, Available at `http://www.jstatsoft.org/v45/i03/`.

[64] S. Van Buuren, H.C. Boshuizen, D.L. Knook, *et al.*, *Multiple imputation of missing blood pressure covariates in survival analysis*, Statistics in Medicine 18 (1999), pp. 681–694.

[65] E. von Elm, D.G. Altman, M. Egger, S. Pocock, P. Gøtzsche, J. Vandenbroucke, and for the STROBE Initiative, *The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies*, Epidemiology 18 (2007), pp. 800–804.

[66] I.R. White, P. Royston, and A.M. Wood, *Multiple imputation using chained equations: Issues and guidance for practice*, Statistics in medicine 30 (2011), pp. 377–399.